The Plant Cell, Vol. 26: 2367–2389, June 2014, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

Meta-Analysis of *Arabidopsis thaliana* Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs[C][W]

Klaas J. van Wijk,[a,1] Giulia Friso,[a] Dirk Walther,[b] and Waltraud X. Schulze[c]

[a] Department of Plant Biology, Cornell University, Ithaca, New York 14850
[b] Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany
[c] Department of Plant Systems Biology, University of Hohenheim, 70593 Stuttgart, Germany

ORCID ID: 0000-0001-9536-0487 (K.J.v.W.)

Protein (de)phosphorylation plays an important role in plants. To provide a robust foundation for subcellular phosphorylation signaling network analysis and kinase-substrate relationships, we performed a meta-analysis of 27 published and unpublished in-house mass spectrometry–based phospho-proteome data sets for *Arabidopsis thaliana* covering a range of processes, (non)photosynthetic tissue types, and cell cultures. This resulted in an assembly of 60,366 phospho-peptides matching to 8141 nonredundant proteins. Filtering the data for quality and consistency generated a set of medium and a set of high confidence phospho-proteins and their assigned phospho-sites. The relation between single and multiphosphorylated peptides is discussed. The distribution of p-proteins across cellular functions and subcellular compartments was determined and showed overrepresentation of protein kinases. Extensive differences in frequency of pY were found between individual studies due to proteomics and mass spectrometry workflows. Interestingly, pY was underrepresented in peroxisomes but overrepresented in mitochondria. Using motif-finding algorithms *motif-x* and MMFPh at high stringency, we identified compartmentalization of phosphorylation motifs likely reflecting localized kinase activity. The filtering of the data assembly improved signal/noise ratio for such motifs. Identified motifs were linked to kinases through (bioinformatic) enrichment analysis. This study also provides insight into the challenges/pitfalls of using large-scale phospho-proteomic data sets to nonexperts.

INTRODUCTION

Phosphorylation is one of the most widespread protein modifications (Lemeer and Heck, 2009) and plays a major role in signal transduction by altering protein activities, protein interactions, or subcellular location (Pawson and Scott, 1997; Seet et al., 2006). Phosphorylation is a reversible modification catalyzed by kinases, which transfer a phosphoryl group from ATP particularly to the hydroxyl group of specific serine, threonine, or tyrosine residues within their target proteins (Champion et al., 2004). However, basic residues arginine (in particular in bacteria; Elsholz et al., 2012) and histidine (in particular for histones and two-component signaling) also can be phosphorylated, but mostly due to technical challenges these are often not considered or identified (Cieśla et al., 2011). Phosphatases are responsible for removing phosphorylated residues from the modified proteins. Plant genomes encode about twice as many kinases compared with mammalian genomes (Zulawski et al., 2013). In the genome of the model plant species *Arabidopsis thaliana*, 1052 protein kinases and 162 phosphatases are found (Wang et al., 2014), indicating the important role of protein phosphorylation in regulating cellular processes in the life cycle of plants.

In the past decade, more than 20 medium- to large-scale mass spectrometry–based phospho-proteome (p-proteome) studies in plants have resulted in the identification of many phospho-proteins (p-proteins) and their phosphorylation sites (p-sites) (reviewed in Nakagami et al., 2012). In these studies, different external stimuli and conditions were used to investigate the p-proteome of whole organs (e.g., root or shoot), cell cultures, or specific subcellular fractions (e.g., plasma membranes) (Benschop et al., 2007; Nühse et al., 2007; Carroll et al., 2008; Sugiyama et al., 2008; Whiteman et al., 2008; Jones et al., 2009; Li et al., 2009; Reiland et al., 2009; Umezawa et al., 2009; Chen et al., 2010; Kline et al., 2010; Nakagami et al., 2010; Reiland et al., 2011; Engelsberger and Schulze, 2012; Lan et al., 2012; Mayank et al., 2012; X. Wang et al., 2012; Umezawa et al., 2013; P. Wang et al., 2013; X. Wang et al., 2013; Wu et al., 2013; Yang et al., 2013; Zhang et al., 2013). A majority of the data is hosted in public databases such as PhosPhAt (Durek et al., 2010) or P3DB (Yao et al., 2012) and accessible through the joint portal of MASCP Gator (Joshi et al., 2011). Several plant p-proteomics studies also included prediction of p-site motifs, based on experimental data using the predictor *motif-x* (Chou

[1] Address correspondence to kv35@cornell.edu.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Klaas J. van Wijk (kv35@cornell.edu).
[C] Some figures in this article are displayed in color online but in black and white in the print edition.
[W] Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.114.125815

and Schwartz, 2011), that are potentially recognized by kinases and phosphatases (Sugiyama et al., 2008; Reiland et al., 2009; Liu et al., 2011; Melo-Braga et al., 2012; X. Wang et al., 2013; Zeng et al., 2014). Experimental and predicted protein p-sites from PhosPhAt have been used to assess conservation of p-sites in single-nucleotide polymorphisms (Riaño-Pachón et al., 2010) or phosphorylation hot spots (Christian et al., 2012). In other (non-plant) organisms, large-scale p-proteomics data sets have been successfully used to analyze kinase-specific phosphorylation (Linding et al., 2008; Xue et al., 2008; Newman et al., 2013), and in combination with protein–protein interaction data (Yachie et al., 2011; Song et al., 2012), this information has served to identify the evolution of protein regulation through protein phosphorylation (Boekhorst et al., 2008; Beltrao et al., 2009; Gnad et al., 2010; Pearlman et al., 2011). On a smaller scale, a cross-species comparison were performed between rice (*Oryza sativa*) and *Arabidopsis* (Nakagami et al., 2010), allowing for a discussion of the role of conserved protein p-sites.

The distribution of predicted *Arabidopsis* kinases across organs, cell types, and subcellular locations is poorly understood. However, it is very likely that specific kinases and/or kinase families show differential subcellular distribution and activity. For instance, even if chloroplasts contain some 3000 predicted proteins, until recently, only four protein kinases have been conclusively demonstrated to localize to plastids (STN7, STN8, CSK, and PKT) (reviewed in Schliebner et al., 2008; Bayer et al., 2012; Schönberg and Baginsky, 2012). However, recently, a set of 17 atypical ABC1 kinases was discovered that appears to localize exclusively to mitochondria and chloroplasts (Lundquist et al., 2012). Substrates of these ABC1 kinases may have specific phosphorylation motifs (p-motifs), which should then be present only in plastid and mitochondria and identification of chloroplast or mitochondrial-enriched p-motifs could help determine kinase-substrate relationships.

Now that a significant number of *Arabidopsis* p-proteomics data sets are publicly available, it is possible to systematically assemble and integrate these with meta-information on experimental conditions and workflows. Such meta-analysis of *Arabidopsis* p-proteomics data will allow curation for p-proteome and p-site quality, and the ability to obtain novel information on p-site motifs and distributions across subcellular compartments and functions. Here, we assembled the most comprehensive set of *Arabidopsis* p-proteins and their assigned p-sites, including metadata with information on plant and sample treatments, p-peptide enrichment, and proteomics workflow from 27 published articles and additional unpublished data. We explored the assembled data set for possible biases such as observations of pY-sites, multiphosphorylated peptides, and phosphopeptides of proteins involved in photosynthesis. After filtering for consistency and quality, we then determined (1) the distribution of p-proteins across functions and subcellular locations and (2) p-motifs for different subcellular locations using two different motif finders. We compare p-motifs with published motifs for *Arabidopsis* and other plants species, such as poplar (*Populus trichocarpa*; Liu et al., 2011), citrus fruit (*Citrus sinensus*; Zeng et al., 2014), and grape (*Vitis vinifera*; Melo-Braga et al., 2012). All p-protein and p-site information used here is available through the PhosPhAt database (http://phosphat.mpimp-golm.mpg.de/), and p-protein sets with their annotations are available through the Plant Proteome Database (PPDB; athttp://ppdb.tc.cornell.edu/).

RESULTS

Assembly of the *Arabidopsis* p-Proteome

To assemble the current p-proteome and their p-sites, we collected the 27 most significant *Arabidopsis* p-proteomics data sets published in the literature (until February 2013), supplemented by in-house unpublished data sets. These data are available through the PhosPhAt database (Heazlewood et al., 2008; Durek et al., 2010). Metadata were extracted from the articles, including plant growth and treatment conditions, sample preparation and enrichment information, and mass spectrometry (MS) acquisition with search methods and cutoff filters. Table 1 provides a summary of the data sets, their metadata, as well as a numerical overview of p-peptides and p-proteins imported into PhosPhAt; more details are provided in Supplemental Data Set 1.

The assembly consists of 60,366 p-peptides including their charge state and calculated number of phosphates (from the parent ion mass) with associated data that were matched (in the respective articles) to 8141 nonredundant proteins and genes (Supplemental Data Sets 2 and 3). These proteins were here annotated for name, function using MapMan (Thimm et al., 2004), and subcellular localization using information from PPDB supplemented with localization information from SUBA3 (http://suba.plantenergy.uwa.edu.au/) (Supplemental Data Sets 1 and 2).

Of the 27 data sets analyzed, four studies used a quadrupole time-of-flight mass spectrometer (MicroMass/Waters), eight used a linear triple quadrupole (LTQ) mass spectrometer, one used an Orbitrap-Fourier transform ion cyclotron resonance (FTICR) mass spectrometer, and the other studies and internal data sets were acquired on a LTQ-Orbitrap (Supplemental Data Set 1). The LTQ has the lowest mass accuracy and the FTICR the highest. About half of the studies used TiO_2 for p-peptide enrichment and the rest employed various types of IMAC enrichment (mostly Fe). Most studies used the search engine Mascot, whereas the LTQ users mostly employed Sequest. The different studies also varied in their application of significance thresholds, using maximum P values of 0.05 or 0.01 (Mascot) or Xcorr>2 or 2.5 (Sequest) and various ion score cutoffs (no cutoff, >30 to 50 ppm) and some used a specific cutoff for post-translational modifications (PTMs). The precursor ion threshold depended on the instrument (3 to 30 ppm for LTQ-Orbitrap, 300 ppm for LTQ, and 30 ppm for quadrupole time-of-flight). Eleven of the studies used cell cultures, and the rest used soil-, plate-, or hydroponically grown tissue, including young seedlings and leaves/rosettes (13x), roots (1x), pollen (1x), or seeds (2x) (Table 1). Thus, cell cultures and photosynthetic leaf material are overrepresented in the metadata set, while undersampling roots, seeds, and the inflorescence. Various types of stress treatments were employed, in particular nutrient stress (nitrogen, sugars, and phosphate), but also hormone treatments (e.g., ethylene and abscisic acid) and the biotic stressor flg22. The different studies showed a range of subcellular sample type enrichment, such as plasma membranes, nuclei, mitochondria, or simply total

Table 1. Summary of the 27 Published p-Proteomics Studies and Unpublished in-House Data Used for This Meta-Analysis

Reference	Tissue	Photosynthetic Tissue	Treatment[a]	Cellular Comp.[b]	No. of Rp-Proteins[c]	No. of Rp-Peptides[d]	% pS[e]	% pT[e]	% pY[e]
Nühse et al. (2004)	Cell culture	No	None	PM	195	386	85	14	1
Wolschin and Weckwerth (2005)	Seed	No	None	Total	10	16	88	13	0
de la Fuente van Bentem et al. (2006)	Cell culture	No	None	cytosol	22	52	98	2	0
Benschop et al. (2007)	Cell culture	No	flg22 or xylanase	PM	431	746	85	15	0
Niitylä et al. (2007)	Seedlings	Yes	Sucrose starvation/sucrose resupply	PM	70	83	74	19	7
Nühse et al. (2007)	Cell culture	No	None	PM	121	189	91	9	0
Carroll et al. (2008)	Cell culture	No	None	Cytosol, mitochondria	70	83	100	0	0
de la Fuente et al. (2008)	Cell culture	No	None	Cytosol	31	31	67	25	8
Sugiyama et al. (2008)	Cell culture	No	None	Total	1275	2577	86	10	4
Whiteman et al. (2008)	Leaves	Yes	None	Tonoplast-enriched	142	179	91	9	0
Jones et al. (2009)	Seedlings	Yes	None	Enriched nuclei	259	303	92	8	0
Li et al. (2009)	Seedlings	Yes	Ethylene or ambient air	Total	186	218	80	18	1
Wang et al. (2009)	Leaves	Yes	Nitrate starvation and resupply	Total	24	28	96	4	0
Reiland et al. (2009)	Rosette and rosette	Yes	End of night or day; none	Total	1429	3829	89	11	1
Ito et al. (2009)	Cell culture	No	None	Mitochondria	73	94	91	9	0
Umezawa et al. (2009)*	Seedlings	Yes	ABA	Total	3	8	80	20	0
Chen et al. (2010)	Cell culture	No	ABA/GA/JA/IAA/kinetin	PM	129	147	92	8	1
Nakagami et al. (2010)	Cell culture	No	None	Total	2152	5061	84	12	4
Kline et al. (2010)*	Seedlings	Yes	Abscisic acid	Total	60	64	89	9	2
Reiland et al. (2011)	Rosette	Yes	None	Soluble and membranes	1607	3905	91	9	0
Engelsberger and Schulze (2012)	Seedlings	Yes	N starvation and resupply/full supply	Cytosol and PM	933	1021	55	32	13
X. Wang et al. (2012)*	Seedlings	Yes	N starvation and resupply	Total protein	18	25	76	24	0
Lan et al. (2012)	Roots	No	Iron deficiency	Total protein	173	239	92	8	1
Meyer et al. (2012)	Seed	No	Seed development	Total protein	170	321	84	11	5
Mayank et al. (2012)	Pollen	No	None	Soluble and membranes	582	975	86	14	0
X. Wang et al. (2013)	Seedlings	Yes	N starvation and resupply	Total protein	2696	5472	79	17	3
Wu et al. (2013)	Seedlings	Yes	Sucrose starvation	Soluble and microsomes	312	465	61	28	11
Schulze lab, internal data	Seedlings/cell culture/leaves/seed	Yes/no	P, N, sucrose starvation, and resupply/full nutrition/anoxia/isoxabene	Cytosol, PM, total	2176	2526	57	31	12

For studies marked with an asterisk, we could only access the p-proteins and p-peptides with assigned p-sites.

[a]ABA, abscisic acid; GA, gibberellic acid; JA, jasmonic acid; IAA, indole-3-acetic acid.

[b]PM, plasma membranes.

[c]Number of nonredundant (NR) p-proteins.

[d]Number of nonredundant (NR) p-peptides.

[e]The percentage of pS, pT, or pY phospho-sites.

membranes or total soluble proteomes, whereas others used total cellular extracts. Finally, the data sets also greatly differed in the number of identified p-proteins and p-peptides. Thus, the metadata set covers many types of biological and technical variables, likely also differing in the false-positive rate of p-peptide identification and p-site assignment.

Quality Control of the p-Proteome Data Sets

Given the wide variation in methodologies between the various studies, we employed several filters to remove possible false-positive p-peptides and/or p-proteins. We generated two data sets by filtering at two levels of stringency (Figure 1). The starting point for our meta-analysis is the set of 60,366 p-peptides with their charge state and calculated number of phosphates from their parent ion masses (# Pi), other experimental information, and their protein match. To identify p-motifs, we extracted the peptide sequences surrounding each phosphorylated residue (pS, pT, or pY) with this p-residue in the central position. We chose to select seven residues directly upstream and seven residues directly downstream for each p-site, resulting in 15-mer sequences. We refer to these as phospho-15-mers, abbreviated as p-15-mers. The set of 60,366 p-peptides represented 8141 proteins and 17,124 different p-15-mers (Figure 1).

For all internal data sets and data sets from the published papers, all matched tandem MS (MS/MS) observations for p-peptides that were reported are included in the metadata set. Except for just three articles (Wolschin and Weckwerth, 2005; de la Fuente van Bentem et al., 2006; Wang et al., 2009), each with low peptide numbers, we found redundant peptides within the extracted data sets from the papers. To reduce likely false-positive observations, we removed proteins observed by just a single p-peptide (MS/MS spectrum). Additionally, we removed p-proteins that were only observed by p-peptides without an assigned p-site since such p-peptides have a higher false-positive rate and are also not informative for finding p-motifs. This first filter removed 4185 p-proteins (Figure 1).

Most p-peptides were predicted to be singly (86%) or doubly (11%) phosphorylated; however, an additional 1095 p-peptides were suggested to have 3 to 10 phosphates (2.5%) or there was no information as to how many phosphates the peptide contained (0.4%) (Figure 2A). The lack of information on the number of phosphates makes the identification of such peptide less certain, since the theoretical mass is then unknown. For various experimental reasons (tight affinity to metal affinity columns, increased ion suppression, and poor MS/MS fragmentation; for discussion, see Engholm-Keller and Larsen, 2013), p-peptides with three or more phosphates are less likely to be identified and/or p-sites within such multiphosphorylated peptides are less likely to be accurately assigned. For these reasons, we assumed that p-peptides with three or more phosphates are less likely to result in reliable information. However, it was reported that pY-containing peptides are

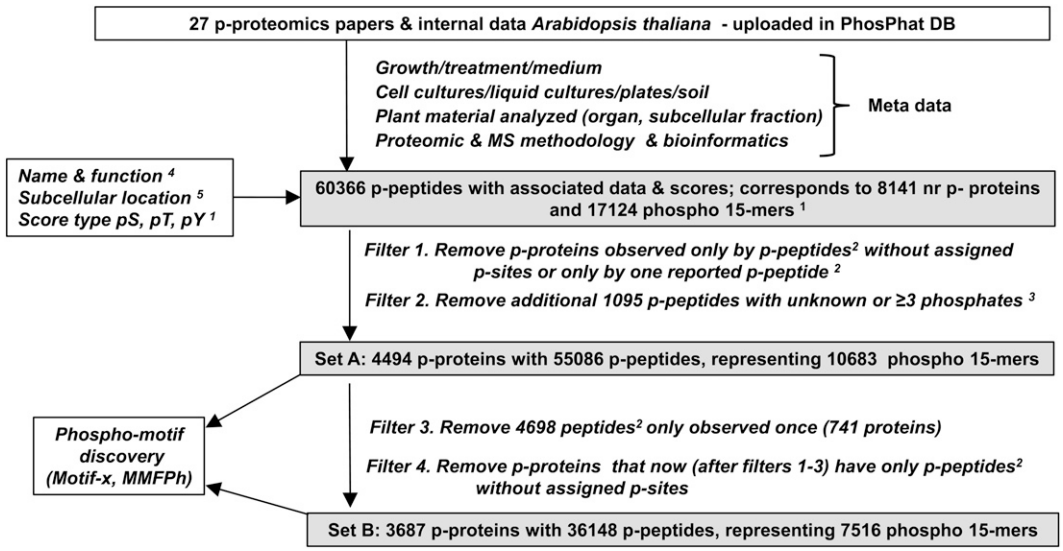


Figure 1. Summary of the Processing Workflow of *Arabidopsis* p-Proteomics Studies.

[1]Phosho 15-mers (p-15-mers) are peptide sequences surrounding the phosphorylated residue (pS, pT, or pY), with this p-residue in the central position (#8); p-15mers can only be obtained if the p-site is assigned. [2]p-peptides refer to reported phosphopeptides extracted from the 27 publications or internal data. In filter 1, one reported p-peptide indicates that for that peptide sequence there was only a single observation, irrespective of charge state, number of phosphates, or other PTMs. For a few publications, only nonredundant p-peptides could be extracted (e.g., only the peptide and matched MS/MS spectra with the highest ion score were reported), whereas for most other publications, and all internal data, all (redundant) peptides could be extracted. Thus, in the former case, even if there were multiple observations for a p-peptide, only one is listed. We treated all p-peptides the same, even if false-positive rates are likely to be different. [3]For some peptides, no information was available about the number of phosphorylations (predicted from the mass shift and/or observed p-sites) within the peptide. [4]Name and function obtained from updated PPDB; functions are based on updated MapMan bin assignments. [5]Subcellular localization from PPDB. If no subcellular localization was available in PPDB, then we used the consensus prediction from SUBA3. However, all plastid assignments are only based on PPDB.

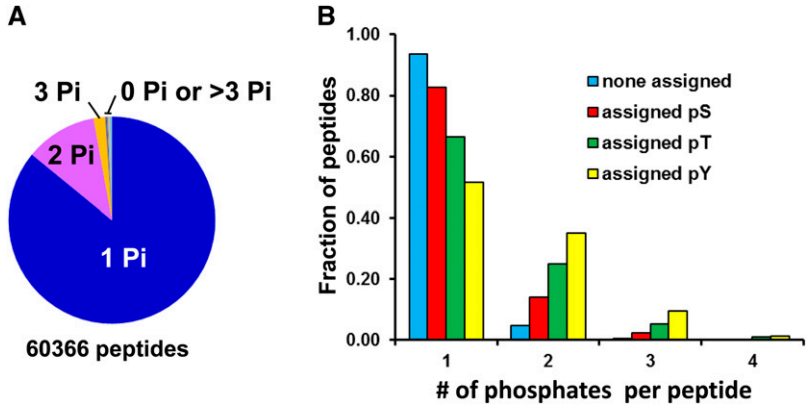


Figure 2. Analysis of Multiphosphorylated Peptides and Assigned p-Sites.

(A) Analysis of all 60,366 p-peptides (Supplemental Table 2) for the number of phosphorylations, showing that most p-peptides are singly (86%) phosphorylated, followed by double (11%) and (1.7%) triple phosphorylation (Pi).

(B) Distribution of peptides with assigned pS, pY, or pT or without any assigned p-site (S) across peptides that are singly, doubly, triply, or quadruply phosphorylated (as calculated from the parent ion mass). This shows that pS-, pT-, and pY-containing peptides are predominantly singly phosphorylated but also that pY is relatively enriched in multiphosphorylated peptides, compared with pT and pS.

frequently (75%) multiphosphorylated (Sugiyama et al., 2008), perhaps explaining why tyrosine phosphorylation is often only observed with very low frequency. To explore this important point, we determined the relationship between p-site and number of phosphorylations per peptide (Figure 2B). In the unfiltered metadata set, 35 and 9.5% of all observed pY-containing p-peptides had two or three phosphates, respectively, whereas 25 and 5.4% pT-containing p-peptides had two or three phosphates, respectively. This was 14 and 2.4% for pS-containing peptides (Figure 2B). Thus, indeed, pY was found more frequently in multiphosphorylated peptides, but not to the extent reported (Sugiyama et al., 2008). Because p-sites observed in triply phosphorylated peptides should also be identified in singly or doubly phosphorylated peptides, and given the lower MS/MS quality of highly phosphorylated peptides, we removed p-peptides with three or more calculated p-sites or an unknown number of p-sites (filter 2). After removing these 1095 p-peptides, 4494 p-proteins with 55,086 p-peptides remained (Supplemental Data Sets 2 and 3), representing 10,683 nonredundant p-15-mers (Supplemental Data Set 4). We name this set A. The metadata set includes substantial redundancy because most experiments were performed with two or more technical or biological replicates. Furthermore, abundant p-proteins are likely to be identified in multiple independent studies. Based on this expected independent discovery, we removed all 4698 peptide sequences (irrespective of charge state, number of phosphorylations, assigned p-sites, and other PTMs) that were only identified once (filter 3), leading to a removal of 741 proteins. Following filters 1 to 3, 12 p-proteins now did not have any p-peptide with assigned p-site left and were therefore also removed (filter 4). The final high confidence set (assigned set B) consisted of 3687 p-proteins identified by 36,148 p-peptides with assigned p-sites (Supplemental Data Sets 2 and 3) and representing 7516 p-15mers (Supplemental Data Set 4).

To illustrate the filters and to demonstrate how multiple independent p-proteomics studies in this meta-proteome analysis can complement each other to identify high frequency p-sites, we selected four p-protein examples (Figures 3 and 4). Details are provided in the legend of Figures 3 and 4. These examples provide insight into the different scenarios encountered when carrying out p-proteomics studies; we believe this will raise awareness as to how to evaluate/appreciate p-proteome information, in particular for the nonexpert reader. The different examples also show how p-sites in triply phosphorylated peptides are sometimes, but not always, supported by singly or doubly phosphorylated peptides.

Data sets A and B were explored with the objective to provide in-depth insight in (1) distribution of p-proteins across cellular functions (based on MapMan), as well as kinase and phosphatase families, (2) distribution of p-proteins across the organelles (plastids, mitochondria, and peroxisomes) and other subcellular compartments, and (3) significant p-motifs within the different subcellular compartments and their respective kinases.

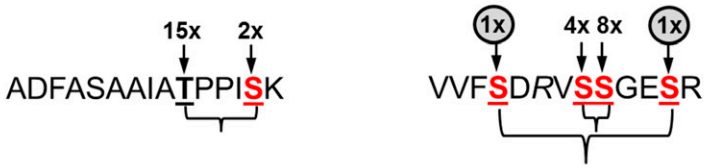
Distribution of p-Proteins across Cellular Functions

Using the MapMan bin system, including updated annotation in PPDB, we determined the distribution of p-proteins across the 34 functions (bins) for data sets A and B. Data sets A and B have essentially the same distribution, showing that the filtering did not bias for protein function (Figure 5A). Proteins with unassigned function represented the largest fraction (32%), followed by protein synthesis and degradation (16%) and RNA metabolism (12%) (Figure 5A). Weighting the distribution to bin size for the whole *Arabidopsis* genome (Figure 5B) showed overrepresentation of proteins involved in signaling (bin 30), cellular organization (bin 31), glycolysis (bin 4), development (bin 33), unknown functions (bin 35), and RNA metabolism (bin 27). Underrepresented were proteins involved in mitochondrial electron transport (bin 9), redox (bin 21), metabolism of cofactor/vitamins (bin 18), tetrapyrroles (bin 19), and amino acids (bin 13) (Figure 5B; note: minimal bin size was 50).

A

gene id	freq (photos.)	freq (non-photo.)	setA	setB	peptide sequence	sequence with PTMs	assigned p-sites	# Pi
AT1G03140		2	x	x	EDELSGGESSDVDADK	EDEL(pS)GGESSDVDADK		1
AT1G03140	1				EDELSGGESSDVDADK	EDEL(pS)GGE(pS)(pS)DVDADK	(pS)(pS)(pS)	3
AT1G03140	1		x		EDELSGGESSDVDADKDMK	EDEL(pS)GGES(pS)DVDADKD(oxM)K	(pS)(pS)	2
AT1G03140	1		x		EDELSGGESSDVDADKDMKR	EDEL(pS)GGESSDVDADKDMKR		1
AT1G03140	1				RAAAAASGGDGKSSGSAPGS SNAATSASSK	RAAAAA(pS)GGDGK(pS)(pS)G(pS) APG(pS)(pS)NAA(t)(s)A(s)(s)K	(pS)(pS)(pS) (pS)(pS)(pS)(pS)	8
AT1G03140	1		x		RLMTFCQR	RL(oxM)(pT)FCQR		1

B



gene id	freq (photos.)	freq (non-photo.)	setA	setB	peptide sequence	sequence with PTMs	assigned p-sites	# Pi	Reference
AT1G01540	9	4	x	x	ADFASAAIATPPISK	ADFASAAIA(pT)PPISK	(pT)	1	Reiland et al (2010, 2011); Wang et al (2013); Benschop et al (2007)
AT1G01540		2	x	x	ADFASAAIATPPISK	ADFASAAIA(pT)PPI(pS)K	(pT)(pS)	2	Nuhse et al (2004); Nakagami et al (2010)
AT1G01540		1	x	x	VVFSDRVSSGESR	VVF(s)DRV(s)(pS)GESR	(pS)	2	Nuhse et al (2004)
AT1G01540		1	x	x	VVFSDRVSSGESR	VVFSDRV(s)(pS)GE(s)R	(pS)	2	Nuhse et al (2004)
AT1G01540		4	x	x	VVFSDRVSSGESR	VVFSDRV(pS)(pS)GESR	(pS)(pS)	2	Nuhse et al (2007); Sugiyama et al (2008); Chen et al (2010); Nakagami et al (2010)
AT1G01540		2	x	x	VVFSDRVSSGESR	VVFSDRVS(pS)GE(pS)R	(pS)(pS)	2	Sugiyama et al (2008); Nakagami et al (2010)
AT1G01540		1			VVFSDRVSSGESR	VVF(pS)DRV(s)(s)GE(pS)R	(pS)(pS)	3	Nuhse et al (2004)

Figure 3. Illustration of the Filters and Multiphosphorylated Peptides in the Metadata Set.

The nonredundant peptide sequences with the P-sites in bold and underlined and the number of observations for each P-site is indicated above the respective sequences. P-sites that are not part of sets A and B are marked by a gray circle. P-sites only observed in photosynthetic tissue or nonphotosynthetic samples (mostly cell cultures) are marked in green and red, respectively (see also inserted table). #Pi refers to the total number of phosphates in a peptide predicted from the parent ion mass.

(A) Nuclear splicing factor Prp18 (At1G03140) was identified by a total of seven p-peptides covering five different, but partially overlapping, peptide sequences. This example illustrates the filters applied to the metadata. Two p-peptides did not qualify for set A or set B because both had more than two predicted phosphates. Three other peptides only qualified for set A but not set B because their peptide sequences were only observed once. Prp18 was identified five times in photosynthetic samples (seedlings and rosettes) in two published studies (Reiland et al., 2011; X. Wang et al., 2013) and one internal data set, and two times in nonphotosynthetic samples (cell cultures) in two studies (Sugiyama et al., 2008; Nakagami et al., 2010). One pS site [L(pS)G], observed in both photosynthetic and nonphotosynthetic samples, was identified five times in three different, but overlapping, peptide sequences, resulting from one or two missed tryptic cleavages in singly, doubly, and triply phosphorylated peptides. Two other pS sites [E(pS)S] and [S(pS)D] were each identified once in, respectively, the triply and doubly phosphorylated peptide. One pT site in a different sequence [RL..QR] was identified in a single observation. One very long peptide (30 amino acids; RA.….SK) and eight predicted phosphates contained six assigned pS sites. Only the pS site in [L(pS)G] passed the filters for inclusion in both sets A and B.

(B) Plasma membrane Tyr kinase RLCK_5 family member (At1g01540) was identified by a total of 24 p-peptides and two different peptide sequences. All but one peptide passed all filters, and they were thus part of sets A and B. This protein was identified nine times in photosynthetic samples (rosettes or seedlings) in three different articles (Reiland et al., 2009, 2011; X. Wang et al., 2013) and 15 times in nonphotosynthetic samples (cell cultures) in six different studies by four different laboratories (Benschop et al., 2007; Nühse et al., 2004, 2007; Sugiyama et al., 2008; Chen et al., 2010; Nakagami et al., 2010). One p-site (pT) was identified 15 times, either in a singly or doubly phosphorylated peptide; this doubly phosphorylated peptide also identified

Distribution of p-Proteins within the Different Subcellular Compartments

Most proteins are localized to specific subcellular compartments within the cell and many compartments have specific sets of kinases and phosphatases. Therefore, it can be postulated that there are specific or enriched motifs within the p-proteomes of the subcellular compartments. To obtain an overview of the distribution of p-proteins across the cell and to search for such enriched p-motifs for specific compartments, we assigned the p-proteins to seven locations as follows: (1) intraplastid, excluding the plastid outer envelope, (2) mitochondrion, (3) peroxisome, (4) nucleus, (5) cytosol, (6) endoplasmic reticulum, Golgi, plasma membrane, cell wall, and vacuolar were placed in one group under the assignment "secretory set," (7) other, in case of proteins with multiple, conflicting, or unknown subcellular locations. The assignment was based on robust, manually curated (i.e., supervised) localizations in PPDB for 1422 proteins, supplemented by unsupervised consensus localization assignment in SUBA3, based on a combination of available experimental evidence and prediction algorithms. In case of plastid localization, only assignment by PPDB was accepted because we recently extensively evaluated all available experimental evidence to assemble a high-quality plastid proteome (Huang et al., 2013). Additional plastid localizations in SUBA3 (Tanz et al., 2013) were mostly based on prediction only, and manual inspection did not provide experimental support for plastid assignment; we therefore assigned them to the location "other." Data sets A (4483 proteins) and B (3679 proteins) had essentially the same subcellular distribution, showing that filtering did not bias for protein location (Figure 5C). Most p-proteins were assigned to the nucleus (35%), followed by the cytosol (23%) and the secretory system (22 to 23%). Eight percent of the p-proteins had no clear location ("other"), whereas plastids, mitochondria, and peroxisomes had 6 to 7% (294 set A/238 set B), 4% (168 set A/135 set B), and 1% (46 set A/39 set B) of the p-proteins, respectively.

Comparison of Seedling/Rosette Samples to Cell Cultures, Roots, and Pollen

About half of the studies concern green tissue (seedlings/rosette) and the other half mostly cell cultures, as well as root, pollen, and seed. Little or no information was generally available with respect to photosynthetic activity in the cell cultures studies, even if most were grown under light/dark cycles, and even if they may contain some chlorophyll based on the light-green color (and assigned heterotrophic). We compared these two sets (seedling/rosette versus cell cultures/root/pollen/seed) for distribution of proteins and peptides across different functions and subcellular locations. Plastid proteins were 50% overrepresented in the seedling/rosette samples, whereas mitochondrial and secretory proteins were 10% underrepresented. Importantly, seedling/rosette samples showed a 10-fold overrepresentation of p-peptides matching to proteins involved in the light and dark reactions of photosynthesis and regulation of the thylakoid electron transport chain (e.g., state transition kinase STN7; AT1G68830) (bin 1) (Figure 5D). Enrichment for photosynthesis was also observed at the p-protein level (3-fold). This showed that the cell cultures were largely nonphotosynthetic and that this meta-analysis can potentially be explored for differences in photosynthetic versus nonphotosynthetic tissue.

Phosphorylation of Protein Kinases and Phosphates

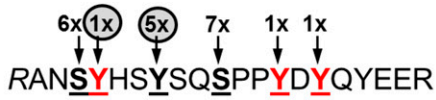
Many protein kinases are themselves regulated by phosphorylation, either through autophosphorylation or by upstream kinases (Bögre et al., 2003; Park et al., 2011; Oh et al., 2012b, 2012c). Because these kinases are central in phosphorylation networks, we evaluated all 8141 identified p-proteins for the presence of a predicted PFAM domain for pKinase or pKinase_Tyr using a threshold of E<0.01 (Supplemental Data Set 3). In total, 532 p-proteins showed significant protein kinase domains (pK, pK_Tyr, or both), representing 6.5% of all identified p-proteins, compared with 3.9% for the complete genome. Classification of these kinases based on Wang et al. (2014) showed that the p-proteome included >100 leucine-rich repeat protein kinase, dozens of cysteine-rich RLKs (receptor-like protein kinases), mitogen-activated protein kinases, calcium-dependent kinases, malectin/receptor-like protein kinases, and many unassigned kinases (Supplemental Data Set 3). On average, about one-third of the members in each kinase family was found as a p-protein in at least one of the filter sets (A, B, or both) (Figure 6A). The highest fractions of phosphorylated family members was found for the families of SnRK1 (66%), SnRK2 (70%), and AGC kinases (61%) (Figure 6B). Among the p-proteins, 64 (out of 162 in the whole genome) are classified as phosphatases (Wang et al., 2014), most of which are in the PP2C family (Supplemental Data Set 3). Others include four BSU1-like (suppressors of leucine-rich repeat receptor-like kinase brassinosteroid insensitive1), Kelch phosphatases, and eight members of the Haloacid dehalogenase–like hydrolases superfamily (Supplemental Data Set 3). Thus, phosphatases were slightly overrepresented in the p-proteome (0.8% compared with 0.6% for the whole genome).

Figure 3. (continued).

twice a pS-site [I(pS)K]. The peptide with sequence [V….R] was observed nine times and identified four pS sites. pS-site [S(pS)G] was identified eight times twice in a doubly phosphorylated peptide, but with the 2nd p-site unassigned, and six times in a double phosphorylated peptides next to another pS site [V(pS)S]. This latter pS-site [V(pS)S] was identified six times in only this doubly phosphorylated peptide. Another pS-site ([F(pS)D] was identified only once in a triply phosphorylated peptide with one more assigned pS site [E(pS)R] and one unassigned p-site. Except for the triply phosphorylated peptide, all peptides passed all filters and were thus present in both sets A and B. This triply phosphorylated peptide has 1 missed cleavage (arginine, marked in italics). Exclusion of the triple p-peptide ignores two pS sites with only a single observation. Coupled observations of pS-sites are indicated by brackets.

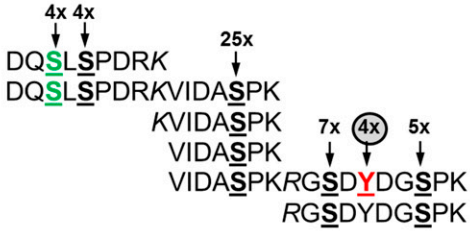
[See online article for color version of this figure.]

A



gene id	freq (photos.)	freq (non-photo.)	setA	setB	peptide sequence	sequence with PTMs	assigned p-sites	# Pi
AT1G08680		1	x	x	RANSYHSYSQSPPYDYQYEER	RAN(s)(y)HSYSQ(pS)PPYDYQYEER	(pS)	2
AT1G08680		1	x	x	RANSYHSYSQSPPYDYQYEER	RAN(pS)YHSYSQSPPYD(pY)QYEER	(pS)(pY)	2
AT1G08680		1	x	x	RANSYHSYSQSPPYDYQYEER	RAN(s)(y)HSYSQSPP(pY)DYQYEER	(pY)	2
AT1G08680		1			RANSYHSYSQSPPYDYQYEER	RAN(pS)(pY)HSYSQ(pS)PPYDYQYEER	(pS)(pY)(pS)	3
AT1G08680		1			RANSYHSYSQSPPYDYQYEER	RAN(s)(y)HS(pY)SQ(pS)PPYDYQYEER	(pY)(pS)	3
AT1G08680	2	2			RANSYHSYSQSPPYDYQYEER	RAN(pS)YHS(pY)SQ(pS)PPYDYQYEER	(pS)(pY)(pS)	3

B



gene id	freq (photos.)	freq (non-photo.)	setA	setB	peptide sequence	sequence with PTMs	assigned p-sites	# Pi	Reference
AT3G53500	1		x	x	DQSLSPDRK	DQ(pS)L(pS)PDRK	(pS)(pS)	2	Wang et al (2013)
AT3G53500		1	x	x	DQSLSPDRK	DQSL(pS)PDRK	(pS)	1	delaFuenteVanBentum (2006)
AT3G53500	1		x	x	DQSLSPDRKVIDASPK	DQ(pS)LSPDRKVIDASPK	(pS)	1	Reiland et al (2011)
AT3G53500	2				DQSLSPDRKVIDASPK	DQ(pS)L(pS)PDRKVIDA(pS)PK	(pS)(pS)(pS)	3	Reiland et al (2009)
AT3G53500	6		x	x	KVIDASPK	KVIDA(pS)PK	(pS)	1	Reiland et al (2009, 2011)
AT3G53500		3	x	x	KVIDASPK	KVIDA(pS)PK	(pS)	1	delaFuenteVanBentum (2006); Wang et al (2009); Mayank et al (2012)
AT3G53500	6		x	x	VIDASPK	VIDA(pS)PK	(pS)	1	Reiland et al (2009, 2011); Wang et al (2013)
AT3G53500		1	x	x	VIDASPK	VIDA(pS)PK	(pS)	1	Wang et al (2009)
AT3G53500	1		x	x	VIDASPKRGSDYDGSPK	VIDA(pS)PKRGSDYDGSPK	(pS)	1	Reiland et al (2011)
AT3G53500		2			VIDASPKRGSDYDGSPK	VIDA(pS)PKRG(pS)DYDG(pS)PK	(pS)(pS)(pS)	3	Sygiyama et al (2008); Nakagami et al (2010)
AT3G53500		2			VIDASPKRGSDYDGSPK	VIDA(pS)PKRG(pS)D(pY)DGSPK	(pS)(pS)(pY)	3	Sygiyama et al (2008); Nakagami et al (2010)
AT3G53500		2			VIDASPKRGSDYDGSPK	VIDA(pS)PKRGSD(pY)DG(pS)PK	(pS)(pY)(pS)	3	Sygiyama et al (2008); Nakagami et al (2010)
AT3G53500	1		x	x	RGSDYDGSPK	RG(pS)DYDGSPK	(pS)	1	Reiland et al (2011)
AT3G53500	1		x		RGSDYDGSPK	RG(s)D(y)DG(s)PK		1	Reiland et al (2011)
AT3G53500	2		x	x	RGSDYDGSPK	RG(pS)DYDG(pS)PK	(pS)(pS)	2	Reiland et al (2009)

Figure 4. Illustration of the Filters and Multiphosphorylated Peptides in the Metadata Set.

The nonredundant peptide sequences with the P-sites in bold and underlined and the number of observations for each P-site is indicated above the respective sequences. P-sites that are not part of sets A and B are marked by a gray circle. P-sites only observed in photosynthetic tissue or nonphotosynthetic samples (mostly cell cultures) are marked in green and red, respectively (see also inserted table). #Pi refers to the total number of phosphates in a peptide predicted from the parent ion mass.

(A) ARF GAP-like zinc finger–containing protein ZIGA4 (At1G08680) was identified by a total of 80 p-peptides and five different peptide sequences. Here, we illustrate the information for one of those peptide sequences, identified two times in photosynthetic samples (seedlings and rosettes) (X. Wang et al., 2013) and seven times in nonphotosynthetic samples (cell cultures) (Nakagami et al., 2010). Peptides were identified as double or triple phosphorylated peptides. In total, two pS sites and four pY sites were assigned. The pS sites were observed six or seven times, whereas only of the pY sites was observed five times, but the others only one time each.

(B) Zn knuckle–containing, serine/arginine-rich protein Splicing Factor 32 (RSZ32) (At3G53500) was identified with 59 p-peptides across 16 different peptide sequences. Here, we summarize the experimental information for six partially overlapping sequences, mapped by 32 p-peptides. This example that meta-analysis allows assembly of a detailed phosphorylation map across overlapping peptide sequences with the different studies complementing and partially supporting each other. Twenty-one p-peptides were identified in photosynthetic samples (rosettes and leaves) from three published studies, and 11 p-peptides were identified in nonphotosynthetic samples (cell cultures and pollen) in five published studies, as indicated. pS in [A(pS)P] was found in each of the five studies and was in total identified 25 times. pS in [Q(pS)L] (pS marked in green) was only found in the photosynthetic samples in all three studies, whereas pY (marked in red) was only found in a triple phosphorylated peptide together with different pairs of pS. This triple phosphorylated peptide has two missed cleavages around the arginine (marked in italics) and when filtering away this triple phosphorylated peptide, all p-sites except pY remained identified. Note that the 3 pS sites in the triple phosphorylated peptide [DQ…..PK] were also identified in single or double phosphorylated peptides, thus providing confidence to these pS site assignments.

[See online article for color version of this figure.]

Biases and Distribution of pS, pT, and pY Phospho-Sites

Serine, threonine, and tyrosine are the main amino acid residues that are targeted for phosphorylation in eukaryotes. The relative frequency of phosphorylation of serine, threonine, and tyrosine is typically of 80 to 85%, 10 to 15%, and 0 to 5%, respectively. The different studies in our meta-analysis showed that serine was indeed the most frequently phosphorylated residue (average 84%) followed by threonine at much lower levels (average 14%). However, there were dramatic differences in the frequency of observation of pY (Table 1), with nearly half of the published studies reporting no tyrosine phosphorylation and others showing a frequency of up to 15% (average 3%). Indeed, nine of the studies did not even search for pY, whereas eight others did search for pY, but found none or a frequency of 0.7% or less (Table 1). These studies concern different types of samples (plasma membranes, total proteins, soluble proteins, and cytosol), different organ/tissue types (pollen, total seedling, and seed), and cell cultures. Higher pY frequencies were observed for a wide variety of samples, including seedlings, cell cultures, and seeds. Hence, it appears that a high pY frequency is neither related to tissue/sample type nor treatment. Identification of pY can be boosted by including the pY immonium ion (216.043 D) from neutral loss in the MS acquisition setup (Steen et al., 2001). Only the studies from the Schulze lab included this immonium ion (during multistage acquisition on the LTQ-Orbitrap; see Wu et al., 2013); indeed, they report the highest frequency for pY (7 to 15%). The large differences in pY frequency between different studies are therefore unlikely to be related to tissue or sample types, or to treatment, but rather to MS workflow and enrichment bias (Bodenmiller et al., 2007).

pS, pT, and pY represented 76, 19, and 5%, respectively, of all p-sites in set A and, similarly, 78, 17, and 5% in set B (Figure 7A). Using the nonredundant p-15-mers for each protein and their assigned subcellular localization, we calculated the distribution of pS, pT, and pY for sets A and B across the seven subcellular locations. Subcellular distribution for total p-sites based on nonredundant p-15-mers was 40% (nucleus), 22% (secretory system), 21% cytosol), 8% (other), 5% (plastids), 3% (mitochondria), and 1% (peroxisomes), both for sets A and B. This was comparable to the subcellular distribution of p-proteins as shown in Figure 5C. We then determined the relative distribution of pS, pT, and pY across the subcellular locations (Figure 7B). Sets A and B showed similar tendencies, with peroxisomes showing low levels of pY (1 to 2%) and mitochondria showing high levels of pY (12 to 19%). Plastids and peroxisomes seemed to be enriched in pT (24 to 27%) compared with the other subcellular locations.

Prediction of p-Site Motifs across Sets A and B

Several algorithms have been developed to uncover linear motifs around protein phosphorylation sites; this includes the most popular *motif-x* algorithm (http://motif-x.med.harvard.edu/) (Schwartz and Gygi, 2005; Chou and Schwartz, 2011), but also more recent ones such as MMFPh (Maximal Motif Finder for Phosphoproteomics data sets; http://www.cs.dartmouth.edu/~cbk/mmfph/) (T. Wang et al., 2012). Following benchmarking of MMFPh against *motif-x* and other algorithms, T. Wang et al. (2012) suggested that MMFPh is a better and more complete motif predictor, in particular for large data sets, such as presented here.

Because nearly all p-motif searches in plants have been performed with *motif-x*, we first used *motif-x* to search for motives across all p-15-mers for pS, pT, and pY for sets A and B. Previously, smaller plant studies used two types of thresholds: a possibility threshold set at $P < 10^{-5}$ or 10^{-6} and either an absolute (3 to 20 peptides) or relative occurrence threshold (1 to 5%) (Supplemental Data Set 6). The authors of *motif-x* suggest a threshold of 10^{-6}, corresponding to a statistical significance of $P = 0.0003$ when using p-15-mers (Chou and Schwartz, 2011). Our meta-analysis is expected to yield a collection of motifs that are the result of multiple kinases, some with a high number of targets and others with far fewer targets. Therefore, we needed to consider different occurrence thresholds, even if lower occurrence thresholds likely result in more false positives. We tested two possibility thresholds (10^{-6} to 10^{-7}) at three relative occurrence thresholds (1, 3, and 5%). Increasing occurrence thresholds resulted in a pronounced decrease of the number of identified motifs, whereas changing the possibility threshold from 10^{-6} to 10^{-7} had little effect, suggesting that indeed 10^{-6} provided sufficient stringency. We therefore further only considered the 10^{-6} threshold (Supplemental Data Set 7). We identified in total 43 pS, 15 pT, and 2 pY motifs across sets A and B and recorded the fold enrichment for each (Supplemental Data Set 7). For motifs found in both sets A and B, the fold enrichment was typically 10 to 20% higher for the more conservative set B, suggesting that we did remove some false-positive p-15-mers by the additional filtering, but the impact was relatively small. Nevertheless, ~50% of the pS motifs were found in both sets, whereas 17 and 4 pS motifs were only found in set A and set B, respectively.

Many of the different pS motifs are related to each other, and the motifs can be grouped into motif types/families, such as the SP type, the SD type, or the DS type (Supplemental Data Set 7). Searching for motifs at the higher occurrence thresholds (3 and 5%) favored less specific motifs and only the three-residue motifs SDxE, RpSxS, and PxSP were found at the higher 3% threshold, whereas 15 motifs with three fixed positions and one with four fixed positions (SDDE) were only found at 1% occurrence. Furthermore, pS motifs with at least 5-fold enrichment were of the SD type (SDDE, 27-fold; SDxD, 10-fold; SDxE, 11-fold), the SP type (SPxR, 11-fold; SPR, 10-fold; SPK, 10-fold; RxSP, 10-fold; GxxSP, 8-fold; SPxK, 8-fold; PxSP, 8-fold), LxRxS (6-fold), RSxS (7-fold), and SExE (6-fold). There were just nine pS motifs (with two fixed residues) at 5% occurrence (SxxxD, SxxD, SxxE, SxxS, SxE, SxS, SP, SxS, and RxxS), with SP showing the highest fold enrichment (4.7-fold). We found 15 motifs for pT, 11 of which were only found in set A, and two (TD and TP) of which were in both sets A and B. Thus, the additional filtering of p-15-mers did reduce the number of pT motifs rather strongly; however, most of these motifs were <2-fold enriched compared with background, making them less significant. Four pT motifs had residues in three fixed positions, namely, TDD, RTxS, and PxTP with high fold enrichment (>5-fold) in set A and SPT with 7-fold enrichment only in set B. All other pT motifs had <2-fold enrichment. Most other motifs only found in set A were variants of R/KxxT, and we suspect these are most likely false

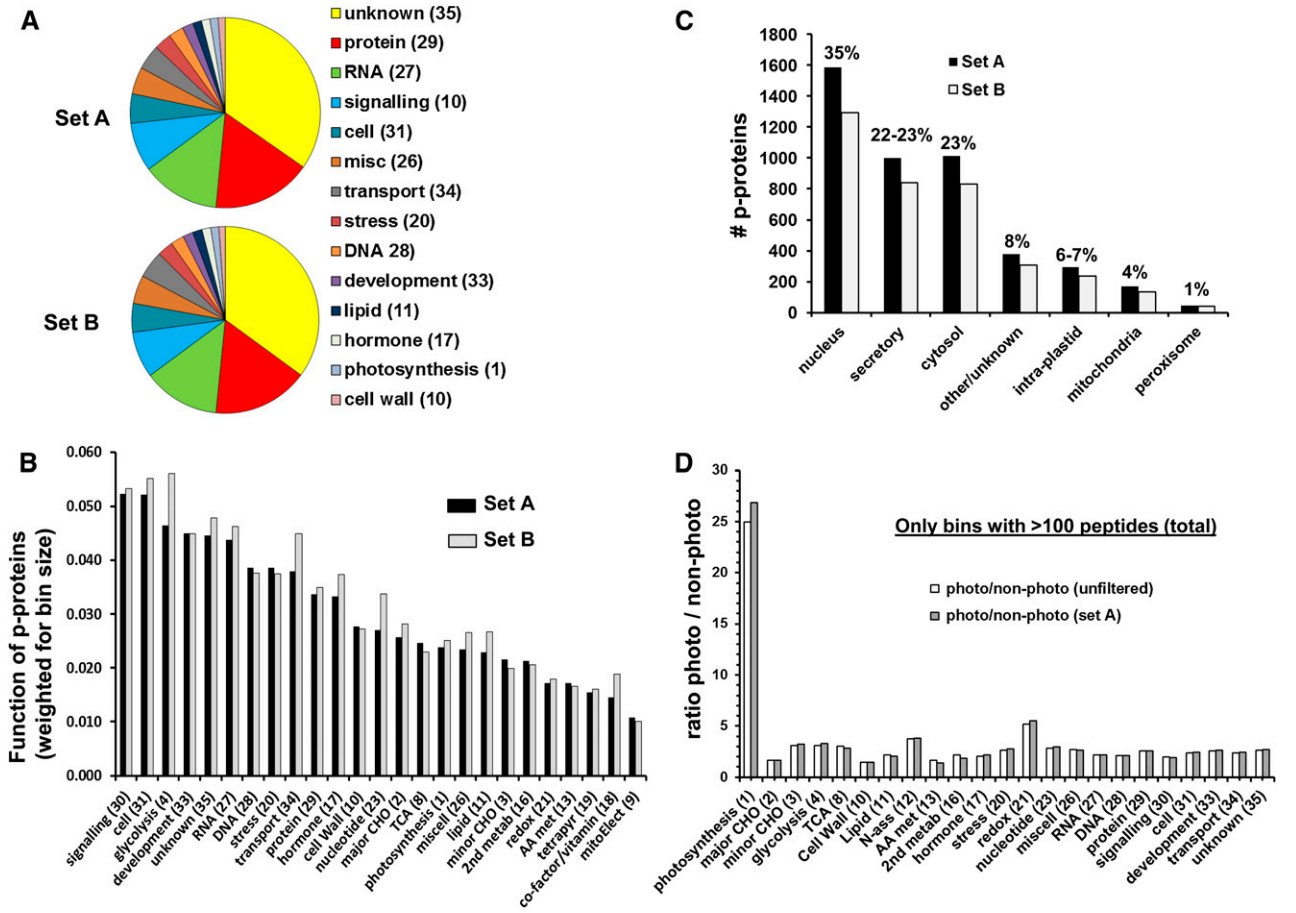


Figure 5. Distribution of p-Proteins across Cellular Functions and Subcellular Compartments.

(A) Distribution of p-proteins in sets A and B across different cellular functions (bins) as listed in PPDB (bins). Only functions (bins) with more than 1% of identified p-proteins are included.

(B) Weighted distribution of p-proteins in sets A and B across different cellular functions (bins) as listed in PPDB (bins). Only functions (bins) with more than 50 proteins are included.

(C) Distribution of p-proteins for set A and set B across the seven subcellular localizations.

(D) Functional distribution of peptides across different cellular functions in studies involving seedlings/rosettes (photosynthetic) versus cell cultures, pollen, roots, and seed (nonphotosynthetic). The diagrams show the ratio in number of peptides per function between the photosynthetic organs and nonphotosynthetic samples in the unfiltered set (squares) or set A (circles) within each MapMan bin.

positive and enriched for the tryptic cleavage site (C-terminal of K/R). In the case of pY, we found two motifs for set A at 1, 3, and 5% occurrence, namely, KY (2.4-fold enriched) and RY (2.3-fold enriched), but none for the more conservative set B.

Using MMFPh at the recommended 10^{-6} default significance threshold in the recommended "complete" mode, as opposed to the "greedy" mode as in *motif-x* (T. Wang et al., 2012), we tested minimum thresholds of 1, 5, and 10% occurrence rate, resulting in 208-47-8 pS motifs for set A and 231-, 44-, and 11 pS motifs for set B (Supplemental Data Set 8). In total, 302 nonredundant pS motifs were identified across sets A and B; thus, 7 times more motifs than for *motif-x*. There was one pS motif with four fixed positions (SDDE), as in *motif-x*, detected at the 1% occurrence threshold in sets A and B. A total of 252 motifs had three fixed positions, and they were all detected at the 1% occurrence threshold. The remaining 49 pS motifs were detected at the 5% and/or 10% threshold and had just two residues at the fixed position; thus, there was no overlap between the motifs detected at 1% and those detected at the 5 and 10% thresholds. However, similar as for *motif-x*, a strong overlap was found between the motifs found in set A and set B at each threshold level. Ten pS motifs were found at the 10% occurrence rate, including motif SP (SxxS, SxxE, SxD, SD, SP, SxS, RxxS, SxxxS, SxxxE, and SxE), five of which were also observed at the highest occurrence rate with *motif-x*. Using the same abundance thresholds (1-5-10%) for pT, we found 13-11-2 pT motifs for set A and 4-4-2 pT motifs for set B, resulting in a total of 19 nonredundant motifs across both sets, compared with 15 motifs with *motif-x* (Supplemental Data Set 8). Eight PT motifs were only found at the 1% threshold and they all had three residues in fixed positions; all other motifs had just two residues in fixed positions (Supplemental Data Set 8). In the case of pY, we found two motifs

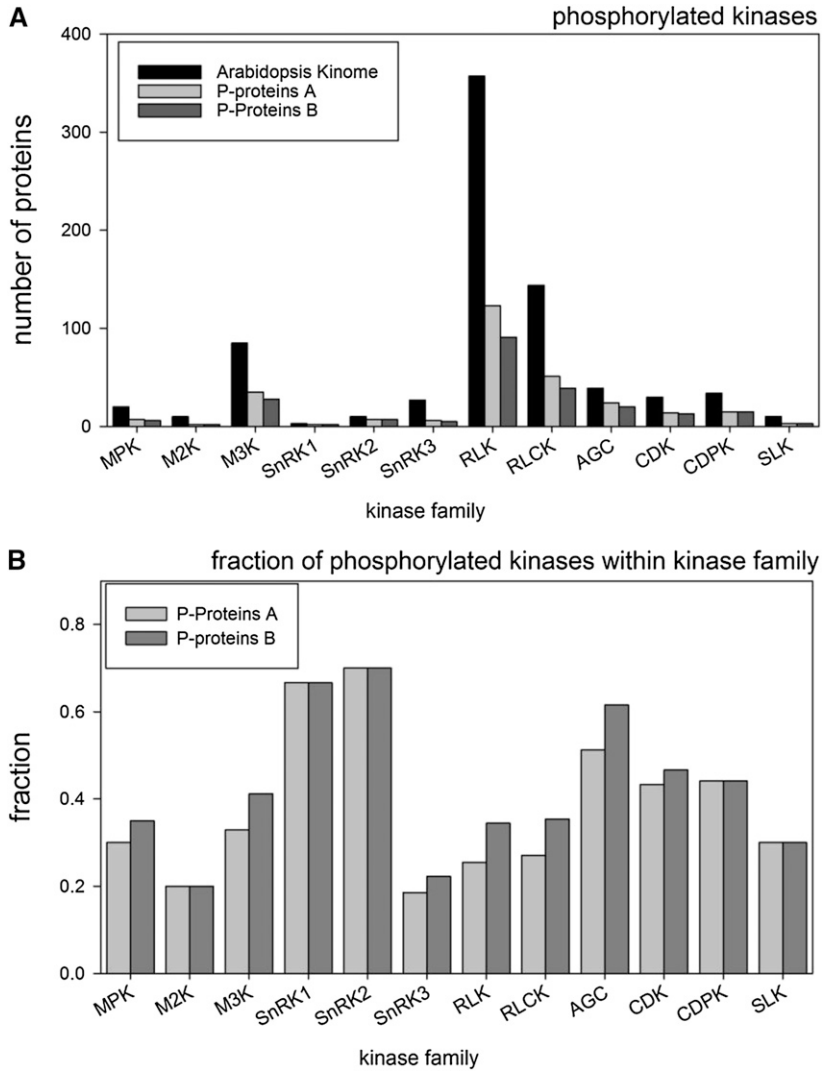


Figure 6. Enrichment Analysis of Kinases in the p-Proteome.

(A) On average, about one-third of the members in each kinase family occurred as a p-protein in at least one of the filter sets (sets A, B, or both).
(B) Highest fractions of phosphorylated family members were found for SnRK1 (66%), SnRK2 (70%), and AGC kinases (61%).

(KY and RY) at 1- 5-10% occurrence rates in set A, but none in set B (Supplemental Data Set 8), similar as for *motif-x*.

Finally, we also checked whether p-15-mers from the unfiltered data sets identified additional motifs, perhaps because we removed triple phosphorylated peptides (filter 2). However, this motif search identified only additional motifs strongly (85%) biased to simple motifs with R or K residues. This R/K enrichment most likely results from the fact that the peptides were generated with trypsin, cleaving the C-terminal of R/K. Therefore, we conclude that the filters did increase the signal/noise ratio for true kinase motifs.

Identification of p-Site Motifs within the Different Subcellular Compartments

To determine if specific p-motifs are enriched in different subcellular locations, we applied both *motif-x* and MMFPh to the seven localization p-15-mers sets for both sets A and B using the same occurrence rates as above. Given that mitochondria and plastids both have a bacterial origin and that more than 100 proteins are dually targeted to both organelles (Carrie and Small, 2013), we also searched the combined mitochondrial and plastid p-15-mer sets in set B. No pY motifs were detected in the localization sets, whereas no pS or pT motifs were detected in mitochondria or peroxisomes.

In the case of *motif-x*, 40 pS motifs and six pT motifs were observed (Supplemental Data Set 7). Nineteen pS motifs were only identified in the nuclear sets; two of these motifs were highly specific (SDDD and SDDE) with >30-fold enrichment compared with the background. Overall, 80% of the pS motifs were detected in the nuclear set, in part likely because the nuclear set contained 40% of all p-15-mers. Interestingly, glycine-enriched motifs (GSxR, SG, and SGP) were exclusively detected in the secretory

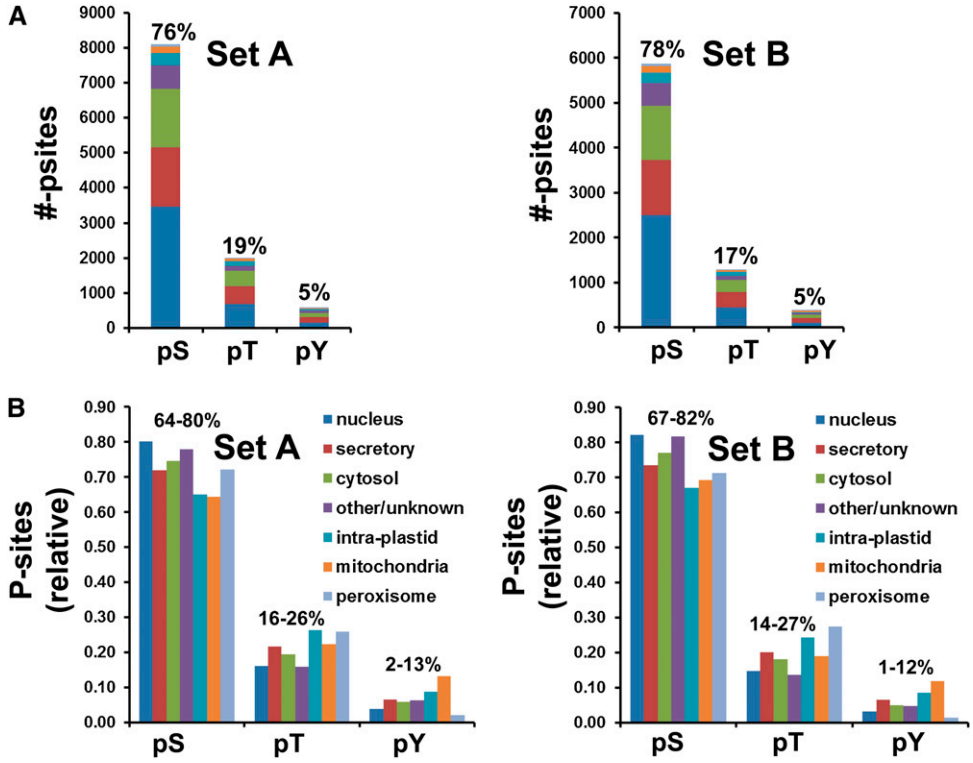


Figure 7. Distribution of pS, pT, and pY 15-mers across Compartments

(A) Absolute number of pS, pT, and pY nonredundant p-15-mers and their distribution across the seven subcellular locations for set A and set B
(B) Relative distribution of pS, pT, and pY nonredundant p-15-mers across the seven subcellular locations for set A and set B. Distribution was weighted for the number of total nonredundant p-15-mers (pS+pT+pY) for each location.

proteome. By contrast, several motifs of the SD type (SD, SDxD, and SDxE), as well as the SP motif, were detected in the cytosol, nucleus, secretome, and plastid/mitochondria, whereas yet other pS motifs were found in nuclei and/or the cytosol (Supplemental Data Set 7). Of the six pT motifs, five (KxxxxxT, TxD, ET, TDD, and KT) were only found in the nuclear set, and in particular motif TDD was highly enriched (12-fold). Motif TP was found in the nucleus, cytosol, secretory system, and plastids with fold enrichment between 3 and 4.2, indicating it is the main pT motif in plants.

Using MMFPH to find motifs in the localization sets at 1-5-10% occurrence levels, we identified a total of 285 pS, 4 pT, and no pY motifs (Supplemental Data Set 8). The highest number of pS motifs was again found in the nuclear set (210-39-13 in set A; 168-36-14 in set B). In the cytosolic set, we identified 11-10-4 motifs in set A and 12-11-9 in set B. In the secretory set, we identified 8-6-4 in set A and 7-7-4 in set B. In "other," we identified 2-2-2 in set A and 5-4-4 in set B; here, the more conservative set yielded more motifs. Finally, we identified either SDxE (at 1% occurrence) or SD (at 5 to 10% occurrence) in plastids, similar as for *motif-x*. Interestingly, when combining the mitochondrial and plastid sets, SDxE, SxDE, and SP motifs were found in both sets A and B and RxxS only in set A. For threonine, we identified motif TxxxE in the nucleus (set A and B) at 1-5-10% and motif TP in both sets A and B at every threshold in the nucleus, cytosol, and secretory sets (Supplemental Data Set 8).

Most Significant p-Motifs across All Data Sets

To systematically compare the results from *motif-x* and MMFPh and identify the most significantly enriched motifs in the various subcellular locations, we merged the results for both search engines (Supplemental Data Set 9). A total of 364 nonredundant pS motifs were observed, with 47 for both search engines, and 10 and 310 only with *motif-x* and MMFPh, respectively. We observed 26 nonredundant pT motifs, and the same KY and RY motifs were found with both motif finders at all thresholds in set A, but never in set B. We used hierarchical clustering to group these pS and pT motifs (Supplemental Figures 1A and 1B).

To identify the most significant motifs, we scored how many times a motif was found in sets A, B, and the subcellular sets by either search engine. In the remainder of Results, we focus on those motifs identified by both engines and/or at least 5 times (out of a possible 80 times). This resulted in 54 pS, 15 pT, and 2 pY motifs, summarized in Table 2. Figures 8A and 8B show a hierarchical tree of these motifs. We recognize nine pS types (clades), namely, glycine-rich, glutamate-rich, DS with basic residues upstream or acidic residues downstream, pS with acid residues, SP motif, pS with basic residues, and serine-rich pS motifs. We recognized seven pT types (clades), namely, glutamate-rich, aspartate-rich, proline-rich (2x), arginine-rich (2x), and lysine-rich pT. The most abundant pS motifs (Table 2) were SP

and RxxS, followed by SD and SDxE, each being identified in all subcellular locations. SxxE, SxE, and SxD were identified in nucleus, cytosol, and, except for SxE, also in plastids/mitochondria. For pT, TP was by far the most abundant motif and was identified in the nucleus, cytosol, and proteins in the "secretory set," but not in plastids/mitochondria. As mentioned earlier, glycine-rich pS motifs SG and SGP were exclusively found in the secreted proteome (marked "S" in Table 2). Moreover, motifs SDDD specifically found by both search engines in the nucleus, and SGP specifically found by both engines in the secreted proteome, were not identified insets A or B. This shows that division of the total p-proteome based on subcellular location can help recover highly specific motifs.

Relating p-Motifs to Proteins and Function

To relate identified motifs to p-proteins and their functions, we linked the most significant motifs (Table 2) back to the p-15-mers and their originating proteins. A significant number of proteins (1695 proteins out of 2752; 62%) showed multiple motifs, which originated most of the time from multiple p-15-mers per protein. However, 2357 of the motif-bearing p-15-mers (41%) contained more than one motif (Supplemental Data Set 10); most of these motifs were variants of each other. Interestingly, multiple p-15-mers per protein, as well as p-15-mers with multiple motifs were particularly found among proteins with functions in transcriptional regulation (MapMan bin 27.3), RNA processing (MapMan bin 27.1), or DNA synthesis and chromatin structure (MapMan bin 28.1). This explains also the high number of motifs identified for nuclear proteins. For example, transcription factor VIP4 (AT5G61150) has 14 p-15-mers covering a total of 20 p-Sites and the motifs SD, SDxD, SxxD, SDDE, SxD, SDxE, SE, SExE, SxxE, SxDxE, SxE, SxS, SP, SPxR, SPR, RxS, and SS. The 20 p-sites found for this protein were located in two large phosphorylation hot spots at the N- and C-terminals of the protein (Christian et al., 2012). It is thus possible that the same p-site can be targeted by different kinases/phosphatases, if the p-site is within two different motifs. For example, SP or SxD can potentially serve as targets for a mitogen-activated protein kinase (MAPK) and RLCK, and RxxxS by another kinase, even if it is the same S.

To address possible kinase-substrate recognition motifs, the p-15-mers bearing specific motifs were tested (using Fischer exact test; P values are indicated below) for enrichment of particular kinase substrates (Supplemental Figure 2). Kinase-substrate relationships that suggested direct interaction were obtained from the kinase target list in PhosPhAt (Zulawski et al., 2013). When only considering direct interactions, for the MAPKs (MPK), the SP motif is the most significant result (P = 0.0008), but SD is a second significant motif (P = 0.0044). The well-known TEY/TDY motif found in MAPKs was not identified as a motif in our analysis, most likely due to low numbers in the data set. The MAPKs themselves were also found to contain peptides with an SP motif. Most of the targets of MAPKs (61%) were of cytoplasmic location. No significant recognition motifs were found for MAPKK (M2K) and MAPKKK (M3K). Among the motifs of the SE, SxE, and SxxE type, we found an over-representation of SnRK2 targets (P = 0.0298, P = 0.066, and P = 0.0019, respectively), RS motifs were particularly significant targets of CDKs (P = 0.029), and 78% of the CDK targets were of nuclear location. RLKs showed a significant preference for SP, SD, and SxS motifs (P = 1E-04, P = 0.031, and P = 0.010, respectively). CDPKs showed a high frequency for SxxD, SD, and SDxE motifs in their targets. An overview of kinase recognition motifs is given in Supplemental Figure 2, in which enrichment P values were converted to probability scores. Probably due to yet limited curated kinase-target information, we could not identify a particular kinase class with preferences for SG, GS, and other motifs. The highest frequency for GS motifs in kinase targets was found for AGC kinases (P = 0.0481); indeed, AGC kinases have a high proportion of plasma membrane proteins in their known targets (41.4%). Among the membrane proteins with a GS or SG motif were NADPH-oxidase RbohD (AT5G47910) and kinases PHOT1 (AT3G45780), CPK9 (AT3G20410), and MAPKKK7 (AT3G13530), which are all known to be regulated by phosphorylation.

DISCUSSION

Benefits of Meta-Analysis of p-Proteomics Studies

Meta-analysis allows the generation of more complete data sets, the recognition of biases or shortcomings of individual studies, the extraction of the most robust data, and the discovery of new patterns and relationships. Because p-proteomes in the individual studies were obtained under many different experimental conditions and from different biological (*Arabidopsis*) materials (e.g., cell lines versus seedlings) and subcellular fractions, this meta-analysis allows identification of p-proteins and p-sites only present/induced under highly specific conditions, as well as basal phosphorylation patterns. Nonphotosynthetic and photosynthetic materials were quite equally represented, but seeds, roots, and the inflorescence were undersampled compared with leaves and cell cultures.

This meta-analysis advances the plant p-proteomics field in multiple ways by providing (1) the most complete and robust set of p-proteins in *Arabidopsis* with their most likely subcellular localization and function, (2) aggregate sets of redundant and nonredundant p-sites for *Arabidopsis*, (3) the most comprehensive and statistically robust p-site motif analysis using two motif finders that employ either a "greedy" mode or a "complete" mode, (4) p-site motifs enriched for specific subcellular localizations, and (5) an enrichment analysis to link kinase families to substrate motifs. This information will aid in the coupling of kinases and their substrates and targeted phosphorylation analysis of specific sets of proteins and provides a strong foundation and reference for analyzing conservation of p-sites across plant species. Detailed p-site data can be viewed in PhosPhAt and the three p-protein data sets (unfiltered and sets A and B) can be extracted with their annotation from PPDB. Importantly, this study also provides insight into the challenges and pitfalls of using large-scale phospho-proteomic data sets to nonexperts, which can help them judge which p-sites are most reliable or most frequent. Here, we will briefly discuss these findings and their implications, integrate this with information from the literature, and provide an outlook for new challenges ahead.

Table 2. Enriched Motifs pS, pT, and pY Found by Motif-x and/or MVFPh for Sets A and B and Subcellular Sets at All Occurrence Thresholds

No. of Motif Merged[a]	Motif	Motif Type	Motif Subtype	Clade No. (Figures 5A and 5B)	Comment	No. of Motifs[b]	*motif-x*[c]	MMFPh[c]	Nucleus[d]	Cytosol[d]	Secreted[d]	Other[d]	Plastid[d]	Plastid and Mito[d]
S-177	.....GS.....	GS	GS	1	Glycine-enriched; highly enriched in secreted; a bit in nucleus	17(12)	A,B,N,S	A,B,N,S	2		10			
S-118	.....SG.....	SG	SG	1	Glycine-enriched; specific to secreted	7(6)	A,S	S			6			
S-74	.....S.G.....	xSx	SxG	1	Glycine-rich S-type	8(4)	A,B,N	A,B	4					
S-13	.....S....E..	xSx	xxSxx	2	Acidic S-type	13(11)	N	A,B,N,C	5	6				
S-2	.....S.....E	xSx	xxSxx	2	Only identified with MMFPh	6(4)		A,B,N	4					
S-7	.....S.....E.	xSx	xxSxx	2	Acidic S-type	7(5)	N	A,B,N	5					
S-150	.....DS.....	DS	DS	3	DS-type	11(8)	A	A,B,N,C,S	4	2	2			
S-269	....R.DS.....	DS	DS	3	DS-subtype enriched in secreted	7(5)	S	A,B,C,S		1	4			
S-160	.....DS.D.....	DS	DSx[D/E]	3	DS-subtype enriched in nucleus	5(3)	N	A,B,N	3					
S-161	.....DS.E.....	DS	DSx[D/E]	3	DS-subtype enriched in nucleus	5(3)	N	A,B,N	3					
S-159	.....DS..E....	DS	DSxx[E/D]	3	DS-subtype; only identified with MMFPh	6(4)		A,B,N,C	2	2				
S-103	.....SDDD....	SD	SDD[D/E]	4	Four-residue SD-type; not identified in total set A and B, clear benefit of making subcellular sets	4(4)	N	N	4					
S-104	.....SDDE....	SD	SDD[D/E]	4	Four-residue SD-type	9(5)	A,B,N	A,B,N,C	4	1				
S-96	.....SD.D....	SD	SDx[D/E]	4	This SD type is very common, target for SLK	15(11)	A,B,N,C,S,P,M	A,B,N,C	4	4	1			2
S-97	.....SD.E....	SD	SDx[D/E]	4	This acidic SD type is very common, target for CDPK	30(24)	A,B,N,C,S,P,M	A,B,N,C,S,P,M	7	4	2		6	5
S-86	.....SD.....	SD	xSDx	4	This SD type is very common. Target for MPK, RLK, CDPK	41(35)	A,B,N,C,S,P,M	A,B,N,C,S,P,M	7	8	8		9	3
S-108	.....SE.....	SE	SE	4	Nuclear and cytosolic, target for SnRK2, AGC	14(11)	A,N,C	A,B,N,C	6	5				
S-113	.....SE.E....	SE	SExE	4	Nuclear and cytosolic	9(5)	A,B,N	A,B,N,C	3	2				
S-54	.....S.D.....	xSx	Sx[D/E]	4	Quite common acidic S-type, target for MPK	21(14)	A,B,N	A,B,N,C,O	9	3				2
S-62	.....S.D.E...	xSx	Sx[D/E]	4	Acidic S-type	5(3)	N	A,B,N	3					
S-64	.....S.DE....	xSx	Sx[D/E]	4	Only identified with MMFPh	6(4)		A,B,N,C,P,M	2	1				1
S-65	.....S.E.....	xSx	Sx[D/E]	4	Quite common acidic S-type, target for SnRK2, RLCK	22(15)	A,B,N,C	A,B,C,N	9	6				
S-33	.....S..D....	xSx	Sxx[D/E]	4	Acidic S-type, target for SnRK1, AGC, CDPK	14(10)	A,C,P,M	A,B,N,C	3	6				1
S-42	.....S..E....	xSx	Sxx[D/E]	4	Acidic S-type enriched in cytosol, target for MPK, SnRK2	23(17)	A,C	A,B,N,C,P,M	4	10				3
S-49	.....S..E.E..	xSx	Sxx[D/E]	4	Acidic S-type	4(2)	N	A,B,N	2					
S-17	.....S....P..	xSx	xxSxx	4	Only identified with MMFPh	5(3)		A,B,N,S	2		1			
S-18	.....S...D...	xSx	xxSxx	4	Acidic S-type	7(2)	B	A,B,N	2					
S-25	.....S...E...	xSx	xxSxx	4	Acidic S-type; only identified with MMFPh	7(4)		A,B,N	4					
S-267	....R..SP.....	SP	SP	5	Basic SP-type, enriched in nucleus	10(6)	A,B,N	A,B,N	6					
S-149	.....SSP.....	pSS	pSS	6	Low frequency motif in nucleus	5(2)	A,B,N	B	2					

(Continued)

Table 2. (continued).

No. of Motif Merged[a]	Motif	Motif Type	Motif Subtype	Clade No. (Figures 5A and 5B)	Comment	No. of Motifs[b]	*motif-x*[c]	MMFPh[c]	Nucleus[d]	Cytosol[d]	Secreted[d]	Other[d]	Plastid[d]	Plastid and Mito[d]
S-120	.....SGP.....	SG	SGP	6	Glycine-enriched; specific to secreted; not identified in total set A and B, clear benefit of making subcellular sets	3(3)	S	S			3			
S-241	....G..SP.....	SP	GxxSP	6	SP type	8(4)	A,B,N	A,B,N	4					
S-212	.....P.SP.....	SP	PxSP	6	SP type	7(2)	A,B	A,B,N	2					
S-121	.....SP.....	SP	SP	6	SP type, extremely common, target for MPK, SnRK2, RLK, AGC, CDK, CDPK, SLK	63(53)	A,B,N,C,O,S,P,M	A,B,N,C,S,O,P,M	11	12	12	9	4	5
S-144	.....SPK.....	SP	SP[K/R]	6	Basic SP-type, enriched in nucleus	8(4)	A,B,N	A,B,N	4					
S-146	.....SPR.....	SP	SP[K/R]	6	Basic SP-type, enriched in nucleus	11(7)	A,B,N,C	A,B,N,C	4	3				
S-139	.....SP.K....	SP	SPx[K/R]	6	Basic SP-type, enriched in nucleus	6(3)	A,B,N	A,N	3					
S-140	.....SP.R....	SP	SPx[K/R]	6	Basic SP-type, enriched in nucleus	10(6)	A,B,N	A,B,N	6					
S-219	.....S.SP.....	SP	SxSP	6	SP subtype, only identified with MMFPh	5(3)		A,B,N,C,O	1	1		1		
S-77	.....S.P.....	xSx	SxP	6	SxP motif; putative 14-3-3 recognition site.	11(8)	A,N,S,O	A,B,N	4		1	3		
S-335	.K.....S.....	xSx	[K/R]xxxxSxx	7	Basic S-type	3(0)	A	A,B						
S-340	.R.....S.....	xSx	[K/R]xxxxSxx	7	Basic S-type	4(1)	A	A,B,N	1					
S-348	K.....S.....	xSx	xxSxx	7	Low-frequency motif in nucleus	5(2)	A	A,B,N	2					
S-183	.....RS.....	RS	RS	8	Does not seem to belong to any specific location, target for AGC, CDK	16(11)	A,B,O	A,B,N,O	2			9		
S-186	.....RS.S.....	RS	RSxS	8	Low-frequency motif in nucleus	6(2)	A,B,N	A,B,N	2					
S-244	....KS.S.....	xSx	[K/R]SxSxx	8	Basic SxS-type	3(1)	A,C	A		1				
S-274	....RS.S.....	xSx	[K/R]SxSxx	8	Basic SxS-type, target for RLK	16(11)	A,B,N,C	A,B,N,C,S,O	3	6	1	1		
S-242	....K..S.....	xSx	[K/R]xxSxx	8	Basic S-type	9(6)	A,N,C,S	A	2	3	1			
S-245	....R..S.....	xSx	[K/R]xxSxx	8	Extremely common motif, target for M2K	60(50)	A,B,N,C,O,S	A,B,C,N,O,S,P,M	11	12	12	12		3
S-319	..L.R..S.....	xSx	xxSxx	8	LxRxS motif, not found in subcellular sets	3(0)	A	A,B						
S-81	.....S.S.....	xSx	pSxS	9	SxS motif	10(6)	B,N,S	B	5		1			
S-216	.....S.S.....	xSx	SxpS	9	SxS motif, target for MPK, RLK, RLCK, CDK	18(10)	A,B	A,B,N,C,S,O	4	2	2	2		
S-300	...S...S.....	xSx	xxSxx	9	Only identified with MMFPh	8(4)		A,B,N	4					
S-31	.....S...S...	xSx	xxSxx	9	SxxS motif not found in subcellular motifs	5(0)	B	A,B						
T-2	.....T....K..	xTx	xTx	1	Basic T motif	4(0)	A	A						
T-3	.....T...E...	xTx	xTx	2	Enriched in nucleus; only identified with MMFPh	6(6)		N	6					
T-5	.....T.E.....	xTx	xTx[E/D]	2	Only identified with MMFPh	7(3)		A,B,C		3				

(Continued)

Table 2. (continued).

No. of Motif Merged[a]	Motif	Motif Type	Motif Subtype	Clade No. (Figures 5A and 5B)	Comment	No. of Motifs[b]	*motif-x*[c]	MMFPh[c]	Nucleus[d]	Cytosol[d]	Secreted[d]	Other[d]	Plastid[d]	Plastid and Mito[d]
T-6	.....TD.....	TD	TD	3	TD only identified with motif-x, likely because MMFPh identified the more specific motif TDD	5(0)	A,B							
T-7	.....TDD.....	TD	TD	3	Low-frequency motif	4(2)	A,N	A	2					
T-1	.....T....D..	xTx	xTx	3	Acidic T motif	3(0)	A	A						
T-4	.....T.D.....	xTx	xTx[E/D]	3	Acidic T motif	2(1)	N	A	1					
T-13	.....PT.....	PT	PT	4	Low-frequency motif	4(0)	B	A,B						
T-16	.....SPT.....	PT	SPT	4	Low-frequency motif	3(0)	B	A,B						
T-15	.....P.TP.....	TP	TP	5	Low-frequency TP motif	3(0)	A	A						
T-8	.....TP.....	TP	TP	5	By far the most common pT motif	45(33)	A,B,N,C,S	A,B,N,C,S	12	11	10			
T-19	....R..T.....	xTx	[R/K]xxT	6	Basic T motif	7(1)	A	A,B,N	1					
T-22	...R...T.....	xTx	xTx	6	Basic T motif	4(0)	A	A						
T-12	.....KT.....	KT	KT	7	Low-frequency motif	6(1)	A,N	A	1					
T-24	.K.....T.....	xTx	xTx	7	Basic T motif; only identified with MMFPh	4(3)		A,N	3					
T-17	....K..T.....	xTx	[R/K]xxT	7	Basic T motif	4(0)	A	A						
Y-1	.....KY.....	KY	KY		Basic T motif	6(0)	A	A,B						
Y-2	.....RY.....	RY	RY		Basic T motif	6(0)	A	A,B						

Numbers of each motif correspond to the numbers in Supplemental Table 7 with more complete information.

[a]Motifs identified by both search engines or at least five times identified across all sets.

[b]Frequency of observation for each motifs across sets A and B and the subsets. The frequency of observation for just the subsets is in parentheses.

[c]A, set A; B, set B; N, nucleus; C, cytosol; S, secretory; O, other; P, plastid; M, mitochondria. This indicated the presence of the respective motif in such (sub)sets.

[d]Number of observations by both search engines in this selected set of motifs.

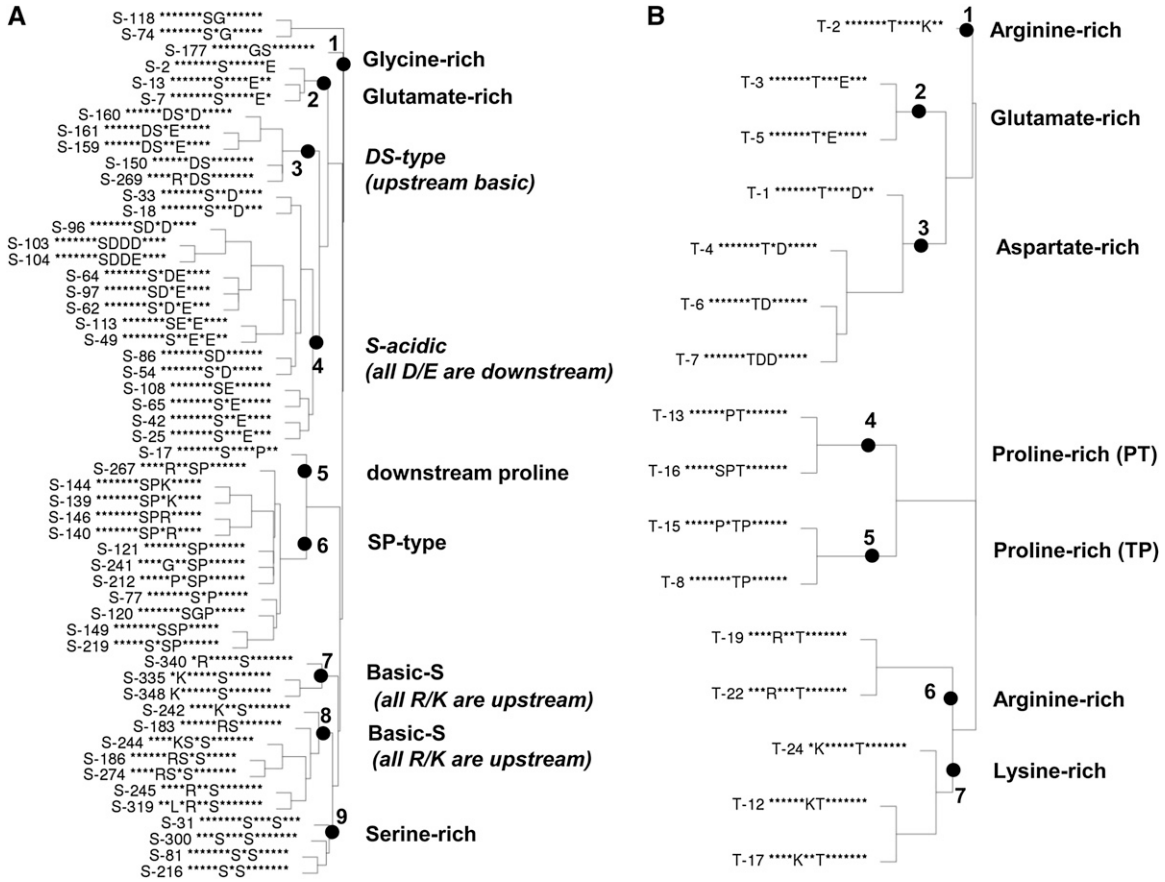


Figure 8. Hierarchical Clustering of the Most Significant pS and pT Motifs Identified by *motif-x* and MMFPh in Sets A and B and the Subcellular Sets.

Hierarchical trees of the most significant 54 pS motifs **(A)** and 16 pT **(B)** motifs, as listed in Table 2.

Filtering of Meta-p-Proteomics Data Removes False Positive p-Peptides and p-Proteins

Starting with the original data from these experimental studies, we significantly reduced the number of p-peptides (by 10 to 40%), p-15-mers (by 47 to 55%), and p-proteins (by 45 to 55%), with the objective of reducing likely false positive information. This improved signal-to-noise ratios is likely to improve the changes to obtain the most biological value and insight, including p-motifs. For a discussion on the challenges of p-site detection and validation in plant p-proteomics studies, see Nakagami et al. (2010). We rejected p-peptides without specific assigned p-sites for the motif analysis and assignment of p-proteins. Because it is not straightforward to measure false-positive rates for p-peptides, p-sites, and p-proteins across this aggregate data set, we applied two different stringencies (resulting in sets A and B) and evaluated both in parallel. The different stringencies did not affect subcellular or functional distribution of p-proteins and pS, pT, or pY p-15-mers. Depending on their objectives, users can choose to compare the uncurated or filtered data sets provided in the supplemental information in conjunction with the Web-based resource PhosPhAt.

Over- and Underrepresentation of Phosphorylation in Cellular Functions and Subcellular Localization

Using subcellular localization assignments based on proteomics and GFP/YPG studies, studies on individual genes, and localization prediction, we were able to generate p-proteome data sets for subcellular compartments. Consequently, this increased the number of known p-proteins in plastids/chloroplasts from ~150 proteins (Reiland et al., 2009, 2011) to nearly 300 p-proteins (294 manually annotated to plastids in set A; excluding outer envelope proteins), which is ~18% of all manually annotated plastid proteins in PPDB. This should facilitate renewed plastid phosphorylation network analysis and linkage between (novel) plastid kinases, such as ABC1Ks (Lundquist et al., 2012) and candidate plastid substrates. According to a recent review (Havelund et al., 2013), 64 mitochondrial proteins (with 103 p-sites) were found to be phosphorylated. This number was increased to 168 proteins (from set A) in this study. This is 50% of the 334 manually annotated mitochondrial proteins (based on experimental data) in PPDB; this high percentage is partially explained by the fact that the mitochondrial annotation for the p-proteome was also based on prediction and weaker identification (from SUBA), but nevertheless demonstrated that phosphorylation is

a frequent and likely important PTM also in mitochondria (see Havelund et al., 2013).

The subcellular distribution of total p-sites based on nonredundant p-15-mers was comparable to the subcellular distribution of p-proteins, but with some overrepresentation in nuclei (40% of all nonredundant p-15-mers), indicating a higher average number of p-sites per protein in this location. This is supported by an enrichment of transcription factors, splicing factors, and DNA binding proteins in "phosphorylation hot spot" proteins (Christian et al., 2012). When weighting against all predicted proteins and their functional annotation, proteins involved in signaling, cellular organization, and development were overrepresented among the phosphorylated proteins, whereas those involved in metabolic functions were generally underrepresented with the exception of glycolysis. Consistent with this overrepresentation, ~6.5% of all p-proteins (8% in set A) were protein kinases, indicating that p-proteins were overrepresented when compared with all predicted kinases (4% of all predicted proteins in the *Arabidopsis* genome). Furthermore, phosphatases were relatively well represented in the p-proteome (0.8% in the p-proteome compared with 0.6% predicted in *Arabidopsis*). In absolute terms, most detected p-proteins have no known function or were involved in protein homeostasis or RNA metabolism.

Tyrosine Phosphorylation

Tyrosine phosphorylation is performed by tyrosine kinases, as well as dual-specificity kinases (DSKs) with S/T p-activity but which can also phosphorylate tyrosines. Bioinformatics analysis of the *Arabidopsis* genome identified 57 candidate DSKs (Rudrabhatla et al., 2006), three putative tyrosine kinases, and 625 tyrosine kinase–like kinases (Martin et al., 2009). In plants, tyrosine phosphorylation was initially deemed insignificant, but large-scale phosphoproteomics in *Arabidopsis* revealed that tyrosine phosphorylation is widespread (Sugiyama et al., 2008; Nakagami et al., 2010, 2012; Ghelis, 2011; Mithoe and Menke, 2011; Oh et al., 2012b, 2012c). Our meta-analysis emphasizes that detection of tyrosine phosphorylation largely depends on proteomics and/or mass spectrometric workflows. In particular for the brassinosteroid signaling pathway, tyrosine (de)phosphorylation has emerged a crucial posttranslational modification and activating pY phosphorylation is well known in MAPKs (Mithoe et al., 2012) and CDPKs (Oh et al., 2012c). In the pY metadata set, we identified four shaggy-like kinases (SK21, SK32, SK41, and SK42), two of which are known components in brassinosteroid signaling, namely, SK21 (AT4G18710, also named BIN2, UCU1, or DWF12) and SK32 (AT4G00720) (Oh et al., 2012b, 2012c; reviewed in Mithoe and Menke, 2011). However, several known pY sites for other components in the brassinosteroid pathway were not in the data set analyzed here. Examples are brassinosteroid (BR) receptor kinase BRI1 (BR-insensitive 1; AT4G39400) and BRI-associated kinase (BAK1; AT5G48380) which are both DSKs that autophosphorylate several of their tyrosines (Kim et al., 2009; Oh et al., 2009, 2012a), whereas BRI1 kinase inhibitor (BKI1) undergoes functionally significant tyrosine phosphorylation (Jaillais et al., 2011). However, BRI1 and several other brassinosteroid signaling components were found to be phosphorylated at serine or threonine (Supplemental Data Set 2). The metadata identified 1185 proteins with tyrosine phosphorylation (Supplemental Data Set 5B) and provide an important basis for further experimentation and assessment of the biological significance of pY in *Arabidopsis* and other plants.

Novel p-Motifs, Their Compartmentalization, and Matching to Functions and Pathways

Multiple studies used p-proteome data to predict p-motifs in *Arabidopsis* (de la Fuente van Bentem et al., 2006; Sugiyama et al., 2008; Reiland et al., 2009; Meyer et al., 2012; Wang et al., 2013) and other plant species, such as poplar (Liu et al., 2011), grape (Melo-Braga et al., 2012), and citrus fruit (Zeng et al., 2014). These predictions were all done using *motif-x* against smaller p-peptide data sets and none involved a meta-analysis. Because the current meta-analysis included all available public data, as well as additional in-house data sets, this meta-p-proteome is much larger and covers many organs, subcellular compartments, experimental growth, and measurement conditions. Consequently, the much larger set of p-15-mers allowed us to do a comprehensive pS and pT motif search using a range of abundance and stringent significance thresholds, as well as motif searches within specific subcellular compartments. Moreover, we used both the popular *motif-x* search engine as well as a novel search engine, MMFPh, previously benchmarked against *motif-x* (T. Wang et al., 2012). Combining the results for both search engines, we then arrived at a set of the most enriched pS and pT motifs, with specific motifs enriched in the nucleus and the secreted proteome (Table 2, Figure 6). The identified motifs include those previously found in smaller studies, as well as many additional motifs. In the case of plastids and mitochondria, we identified a number of enriched pS motifs, in particular SP, RxxS, Sx[D/E], Sxx[D/E], and SDx[D/E], but they are shared with proteins outside of these organelles, likely indicating that these phosphorylations are performed by kinase families located within, as well as outside these organelles. A good example is the casein kinase family (CK1 and CK2) known to have multiple functions and locations in *Arabidopsis* (Lee, 2009; Türkeri et al., 2012; Mulekar and Huq, 2013). No motifs were found for the peroxisomal proteome, most likely because there were insufficient p-proteins assigned to this organelle (70 in set A). Proteins collectively assigned to the "secreted proteome" set (i.e., with signal peptides and/or located in microsomes or vacuoles) showed two specific glycine-rich motifs (SG and SGP) not detected in other compartments. These motifs belonged to 15 proteins, most of which were annotated to the plasma membrane with several belonging to CDPK or PP2C families (Supplemental Data Set 10). Even if the nonredundant p-15-mer set contained 582 (set A) or 365 (Set B) unique pY-mers, this was insufficient to detect pY motifs in the subcellular sets and only KY and RY motifs were detected in sets A and B at all occurrence threshold levels (P value $<10^{-6}$ and 1 to 10% occurrence rate). In a previous study (Sugiyama et al., 2008) (included in this meta-analysis), 11 specific pY motifs or nine degenerate pY motifs were reported based on 95 nonredundant pY-mers using *motif-x*, but no minimal threshold P value was applied or reported. Only one (pYR) of the two motifs (pYK and

pYR) that we identified was also observed in this article (2-fold enriched at 15% occurrence rate [14/95]). Experimental testing to determine the significance of the reported pY motifs is overdue.

The most specific motifs that we detected were SDDD and SDDE; to our knowledge, the kinases recognizing these motifs are not yet known. One of the largest single p-proteomic studies in plants (X. Wang et al., 2013) identified several pS motifs (at 1% occurrence level only), several of which overlap with motifs found here. The SP motif was identified as one of the most frequent, and it seems to be targeted by a number of kinases (Supplemental Figure 2). Motifs SF and RxxSF identified by X. Wang et al. (2013) were not apparent in our study. We believe that our meta-analysis shows that removal of likely false positives and data set size reduction according to subcellular location, in combination with robust evaluation of occurrence thresholds, is important to identify biologically significant motifs.

In addition to linear sequence motifs as determined here, the three-dimensional structure around p-sites may serve as kinase recognition signals. Indeed, a small increase in prediction accuracy was achieved when considering also the spatial neighborhood in addition to sequential proximity (Durek et al., 2009). However, the gain was relatively small, possibly because p-sites have a demonstrated tendency to preferentially occur in unstructured regions (Iakoucheva et al., 2004). The assignment of kinase-cognate motif pairings remains thus a major challenge.

Conclusions and Outlook

This meta-analysis of *Arabidopsis* p-proteomics studies provides a comprehensive repertoire of p-proteins, their p-sites, and p-motifs for pS, pT, and pY. The surprising amount of pY sites provides an important tool to accelerate the significance of pY signaling. This study will aid in the coupling of kinases and their substrates and facilitate targeted phosphorylation analysis of specific sets of proteins. Finally, it provides a strong foundation and reference for analyzing conservation of p-sites across plant species.

METHODS

Data Collection, Extraction, Filtering, and Generation of p-Proteome and p-15-mers Sets

Twenty-seven of the most significant *Arabidopsis thaliana* p-proteomics data sets published in the literature were collected and supplemented with in-house unpublished data sets (Supplemental Data Set 1). These data sets are also uploaded into the PhosPhAt database (Heazlewood et al., 2008; Durek et al., 2010; Zulawski et al., 2013). Meta-data were extracted from the original articles, including plant growth and treatment conditions, sample preparation and enrichment information, and MS acquisition with search methods and cutoff filters. P-proteins and their associated p-peptides were filtered for frequency and quality, resulting in set A and the more conservative set B. To identify p-motifs, peptide sequences surrounding each phosphorylated residue (pS, pT, or pY) were extracted, with this p-residue in the central position. Seven residues directly upstream and seven residues directly downstream of each p-site were selected, resulting in p-15-mer sequences. These are referred to as p-15mers.

Prediction of p-Site Motifs Using Motif-x and MMFPh

To predict p-site motifs, search engines *motif-x* (Chou and Schwartz, 2011) and MMFPh (T. Wang et al., 2012) were used to analyze p-15mer sets A and B. *Motif-x* was used online (http://motif-x.med.harvard.edu/), whereas MMFPh was run locally in the "complete" mode (not "greedy" mode) using the downloaded stand-alone package. Both *Motif-x* and MMFPh were run at the 10^{-6} significance threshold, with the following parameters: "pre-aligned," central S, T, or Y, width 15, occurrences 1-3-5% (given as absolute numbers) and in the case of MMFPh also lbound and hbound 0.05 and debug_mode = FALSE. The background databases for *motif-x* and MMFPh were in both cases all genes in the TAIR10 *Arabidopsis* database, but using all possible gene models in the case of *motif-x* (39,677 proteins) and using one representative protein sequence model per gene in case of MMFPh (27,416 proteins). MMFPh was run on an in-house server where we could select one gene model for every gene, rather than all gene models. The extra sequences are variants of sequences of the TAIR10 set; in many cases not showing differences at the protein level. We believe that it is slightly better to select only one model for each gene, but comparing both databases for MMFPh showed virtually identical results, as expected.

Hierarchical Clustering of Motifs

Motif comparisons were performed using the stringDist routine from the Biostrings R package (http://www.bioconductor.org/packages/2.14/bioc/html/Biostrings.html). Blosum62 exchange sores were used. The exchange score for motif positions associated with no specific amino acid residue type ("*" character) to any other type was set to −1. Clustering was set to average linkage.

Annotation for Function and Subcellular Localization

P-proteins were annotated for name, function (MapMan; Thimm et al., 2004), and, if possible, subcellular localization using information from PPDB (http://ppdb.tc.cornell.edu/) and Huang et al. (2013), supplemented with localization information from SUBA3 (http://suba.plantenergy.uwa.edu.au/). P-proteins were assigned to seven subcellular locations as follows: (1) intraplastid, excluding the plastid outer envelope, (2) mitochondrion, (3) peroxisome, (4) nucleus, (5) cytosol, (6) endoplasmic reticulum, Golgi, plasma membrane, cell wall, and vacuolar assigned "secretory set," (7) other, in case of proteins with multiple, conflicting, or unknown subcellular locations. In case of plastid localization, only assignment by PPDB and Huang et al. (2013) was used, and additional plastid localizations in SUBA3 were therefore assigned to the location "other." Protein kinase targets were obtained from PhosPhAt kinase-target search (Zulawski et al., 2013), and only those kinase-target relationships describing phosphorylation, autophosphorylation, activation, and interaction were considered.

Annotation and Data Display in PPDB and PhosPhat

Detailed phosphorylation information with associated spectral information is available in PhosPhAt (http://phosphat.mpimp-golm.mpg.de/). P-protein data sets (unfiltered, filtered [set A], and conservative filter set B) can also be extracted from PPDB with many associated features through selection of output parameters. The PPDB is hyperlinked at the *Arabidopsis* gene accession level to the PhosPhAt database and the *Arabidopsis* proteome aggregator database MASCP Gator (Joshi et al., 2011) at http://gator.masc-proteomics.org/. In PhosPhAt, motifs identified as a result of this work are available within the motif search function and are highlighted in the sequence view.

Accession Numbers

STATE TRANISTION KINASE7 (STN7), AT1G68830; VERNALIZATION INDEPENDENCE4 (VIP4), AT5G61150; SHAGGY-LIKE KINASE21, BIN2, UCU1, or DWF12, AT4G18710; SHAGGY-LIKE KINASE32 (SK32),

AT4G00720; BR RECEPTOR KINASE1 (BRI1), AT4G39400; BRII-ASSOCIATED KINASE1 (BAK1), AT5G48380; ARF GAP-LIKE ZINC FINGER-CONTAINING PROTEIN (ZIGA4), At1G08680; Zn-knuckle containing, serine/arginine-rich protein splicing factor 32 (RSZ32), At3G53500; PLASMA MEMBRANE TYR KINASE RLCK_5 FAMILY MEMBER, At1G01540; and NUCLEAR SPLICING FACTOR PRP18, At1G03140.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Hierarchical Clustering of All 364 pS and 26 pT Motifs Identified in This Study.

Supplemental Figure 2. Kinase Recognition Motifs for Different Kinase Families in *Arabidopsis*.

The following materials have been deposited in the DRYAD repository under accession number http://dx.doi.org/10.5061/dryad.sb669.

Supplemental Data Set 1. Detailed Overview of the 27 Published p-Proteomics Studies and Unpublished in-House Data with Their Respective Metadata.

Supplemental Data Set 2. The Complete Unfiltered Set of 60,366 p-Peptides with Matched Protein ID, Their Metadata, p-15-mers, Annotation from PPDB, SUBA3 Consensus Prediction, and Assignment to One of Seven Locations.

Supplemental Data Set 3. Nonredundant *Arabidopsis* p-Proteins before Filtering (8141 Proteins) or after Filters 1 and 2 (Set A, 4494 proteins) and after Filters 1 to 4 (set B, 3687 Proteins) with Their Annotations.

Supplemental Data Set 4. Nonredundant p-15-mers Prior to Filtering and for Sets A and B.

Supplemental Data Set 5. Analysis of pY Peptides and pY Proteins.

Supplemental Data Set 6. Published Plant p-Motifs in Various Plant Species Based on *motif-x* Searches against p-Proteomics Data.

Supplemental Data Set 7. P-Motifs for pS, pT, and pY and Their Fold-Enrichment in Sets A and B, and the Localization of p-15-mer Sets Using *motif-x* at the 10^{-6} Threshold and 1, 3, and 5% Occurrence Rates.

Supplemental Data Set 8. P-Motifs for pS, pT, and pY in Sets A and B and the Localization of p-15-mer Sets Using MMFPh at the 10^{-6} Threshold and 1, 5, and 10% Occurrence Rates.

Supplemental Data Set 9. Motifs for pS, pT, and pY Found by *motif-x* and MMFPh for Sets A and B and Subcellular Sets at All Occurrence Thresholds

Supplemental Data Set 10. P-Proteins with Their p-15-mers and Their Most Significant Motifs (from Table 2).

ACKNOWLEDGMENTS

This study was supported by an Alexander von Humboldt Research award (http://www.humboldt-foundation.de) and National Science Foundation Grant IOS-1127017 to K.J.v.W.

AUTHOR CONTRIBUTIONS

K.J.v.W. and W.X.S. conceived the study and wrote the article. G.F., D.W., W.X.S., and K.J.v.W. each carried out various aspects of the motif searches and theoretical analyses.

Received March 27, 2014; revised March 27, 2014; accepted May 9, 2014; published June 3, 2014.

REFERENCES

Bayer, R.G., Stael, S., Rocha, A.G., Mair, A., Vothknecht, U.C., and Teige, M. (2012). Chloroplast-localized protein kinases: a step forward towards a complete inventory. J. Exp. Bot. **63:** 1713–1723.

Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., Burlingame, A.L., and Krogan, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. **7:** e1000134.

Benschop, J.J., Mohammed, S., O'Flaherty, M., Heck, A.J., Slijper, M., and Menke, F.L. (2007). Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis. Mol. Cell. Proteomics **6:** 1198–1214.

Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B., and Aebersold, R. (2007). Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat. Methods **4:** 231–237.

Boekhorst, J., van Breukelen, B., and Heck, A., Jr., and Snel, B. (2008). Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. Genome Biol. **9:** R144.

Bögre, L., Okrész, L., Henriques, R., and Anthony, R.G. (2003). Growth signalling pathways in Arabidopsis and the AGC protein kinases. Trends Plant Sci. **8:** 424–431.

Carrie, C., and Small, I. (2013). A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. Biochim. Biophys. Acta **1833:** 253–259.

Carroll, A.J., Heazlewood, J.L., Ito, J., and Millar, A.H. (2008). Analysis of the Arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification. Mol. Cell. Proteomics **7:** 347–369.

Champion, A., Kreis, M., Mockaitis, K., Picaud, A., and Henry, Y. (2004). Arabidopsis kinome: after the casting. Funct. Integr. Genomics **4:** 163–187.

Chen, Y., Hoehenwarter, W., and Weckwerth, W. (2010). Comparative analysis of phytohormone-responsive phosphoproteins in Arabidopsis thaliana using TiO2-phosphopeptide enrichment and mass accuracy precursor alignment. Plant J. **63:** 1–17.

Chou, M.F., and Schwartz, D. (2011). Biological sequence motif discovery using motif-x. Curr. Protoc. Bioinformatics **13:** 15–24.

Christian, J.O., Braginets, R., Schulze, W.X., and Walther, D. (2012). Characterization and prediction of protein phosphorylation hotspots in *Arabidopsis thaliana*. Front. Plant Sci. **3:** 207.

Cieśla, J., Frączyk, T., and Rode, W. (2011). Phosphorylation of basic amino acid residues in proteins: important but easily missed. Acta Biochim. Pol. **58:** 137–148.

de la Fuente van Bentem, S., Anrather, D., Roitinger, E., Djamei, A., Hufnagl, T., Barta, A., Csaszar, E., Dohnal, I., Lecourieux, D., and Hirt, H. (2006). Phosphoproteomics reveals extensive in vivo phosphorylation of Arabidopsis proteins involved in RNA metabolism. Nucleic Acids Res. **34:** 3267–3278.

Durek, P., Schmidt, R., Heazlewood, J.L., Jones, A., MacLean, D., Nagel, A., Kersten, B., and Schulze, W.X. (2010). PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update. Nucleic Acids Res. **38:** D828–D834.

Durek, P., Schudoma, C., Weckwerth, W., Selbig, J., and Walther, D. (2009). Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. BMC Bioinformatics **10:** 117.

Elsholz, A.K., Turgay, K., Michalik, S., Hessling, B., Gronau, K., Oertel, D., Mäder, U., Bernhardt, J., Becher, D., Hecker, M., and Gerth, U. (2012). Global impact of protein arginine phosphorylation on the physiology of *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **109:** 7451–7456.

Engelsberger, W.R., and Schulze, W.X. (2012). Nitrate and ammonium lead to distinct global dynamic phosphorylation patterns when resupplied to nitrogen-starved Arabidopsis seedlings. Plant J. **69:** 978–995.

Engholm-Keller, K., and Larsen, M.R. (2013). Technologies and challenges in large-scale phosphoproteomics. Proteomics **13:** 910–931.

Ghelis, T. (2011). Signal processing by protein tyrosine phosphorylation in plants. Plant Signal. Behav. **6:** 942–951.

Gnad, F., Forner, F., Zielinska, D.F., Birney, E., Gunawardena, J., and Mann, M. (2010). Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. Mol. Cell. Proteomics **9:** 2642–2653.

Havelund, J.F., Thelen, J.J., and Møller, I.M. (2013). Biochemistry, proteomics, and phosphoproteomics of plant mitochondria from non-photosynthetic cells. Front. Plant Sci. **4:** 51.

Heazlewood, J.L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., and Schulze, W.X. (2008). PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. Nucleic Acids Res. **36:** D1015–D1021.

Huang, M., Friso, G., Nishimura, K., Qu, X., Olinares, P.D., Majeran, W., Sun, Q., and van Wijk, K.J. (2013). Construction of plastid reference proteomes for maize and Arabidopsis and evaluation of their orthologous relationships; the concept of orthoproteomics. J. Proteome Res. **12:** 491–504.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. **32:** 1037–1049.

Ito, J., Taylor, N.L., Castleden, I., Weckwerth, W., Millar, A.H., and Heazlewood, J.L. (2009). A survey of the *Arabidopsis thaliana* mitochondrial phosphoproteome. Proteomics **9:** 4229–4240.

Jaillais, Y., Hothorn, M., Belkhadir, Y., Dabi, T., Nimchuk, Z.L., Meyerowitz, E.M., and Chory, J. (2011). Tyrosine phosphorylation controls brassinosteroid receptor activation by triggering membrane release of its kinase inhibitor. Genes Dev. **25:** 232–237.

Jones, A.M., MacLean, D., Studholme, D.J., Serna-Sanz, A., Andreasson, E., Rathjen, J.P., and Peck, S.C. (2009). Phosphoproteomic analysis of nuclei-enriched fractions from Arabidopsis thaliana. J. Proteomics **72:** 439–451.

Joshi, H.J., et al. (2011). MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. Plant Physiol. **155:** 259–270.

Kim, T.W., Guan, S., Sun, Y., Deng, Z., Tang, W., Shang, J.X., Sun, Y., Burlingame, A.L., and Wang, Z.Y. (2009). Brassinosteroid signal transduction from cell-surface receptor kinases to nuclear transcription factors. Nat. Cell Biol. **11:** 1254–1260.

Kline, K.G., Barrett-Wilt, G.A., and Sussman, M.R. (2010). In planta changes in protein phosphorylation induced by the plant hormone abscisic acid. Proc. Natl. Acad. Sci. USA **107:** 15986–15991.

Lan, P., Li, W., and Schmidt, W. (2012). Complementary proteome and transcriptome profiling in phosphate-deficient Arabidopsis roots reveals multiple levels of gene regulation. Mol. Cell. Proteomics **11:** 1156–1166.

Lee, J.Y. (2009). Versatile casein kinase 1: multiple locations and functions. Plant Signal. Behav. **4:** 652–654.

Lemeer, S., and Heck, A.J. (2009). The phosphoproteomics data explosion. Curr. Opin. Chem. Biol. **13:** 414–420.

Li, H., Wong, W.S., Zhu, L., Guo, H.W., Ecker, J., and Li, N. (2009). Phosphoproteomic analysis of ethylene-regulated protein phosphorylation in etiolated seedlings of Arabidopsis mutant ein2 using two-dimensional separations coupled with a hybrid quadrupole time-of-flight mass spectrometer. Proteomics **9:** 1646–1661.

Linding, R., Jensen, L.J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M.B., and Pawson, T. (2008). NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res. **36:** D695–D699.

Liu, C.C., Liu, C.F., Wang, H.X., Shen, Z.Y., Yang, C.P., and Wei, Z.G. (2011). Identification and analysis of phosphorylation status of proteins in dormant terminal buds of poplar. BMC Plant Biol. **11:** 158.

Lundquist, P.K., Davis, J.I., and van Wijk, K.J. (2012). ABC1K atypical kinases in plants: filling the organellar kinase void. Trends Plant Sci. **17:** 546–555.

Martin, D.M., Miranda-Saavedra, D., and Barton, G.J. (2009). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. Nucleic Acids Res. **37:** D244–D250.

Mayank, P., Grossman, J., Wuest, S., Boisson-Dernier, A., Roschitzki, B., Nanni, P., Nühse, T., and Grossniklaus, U. (2012). Characterization of the phosphoproteome of mature Arabidopsis pollen. Plant J. **72:** 89–101.

Melo-Braga, M.N., Verano-Braga, T., León, I.R., Antonacci, D., Nogueira, F.C., Thelen, J.J., Larsen, M.R., and Palmisano, G. (2012). Modulation of protein phosphorylation, N-glycosylation and Lys-acetylation in grape (*Vitis vinifera*) mesocarp and exocarp owing to *Lobesia botrana* infection. Mol. Cell. Proteomics **11:** 945–956.

Meyer, L.J., Gao, J., Xu, D., and Thelen, J.J. (2012). Phosphoproteomic analysis of seed maturation in Arabidopsis, rapeseed, and soybean. Plant Physiol. **159:** 517–528.

Mithoe, S.C., and Menke, F.L. (2011). Phosphoproteomics perspective on plant signal transduction and tyrosine phosphorylation. Phytochemistry **72:** 997–1006.

Mithoe, S.C., Boersema, P.J., Berke, L., Snel, B., Heck, A.J., and Menke, F.L. (2012). Targeted quantitative phosphoproteomics approach for the detection of phospho-tyrosine signaling in plants. J. Proteome Res. **11:** 438–448.

Mulekar, J.J., and Huq, E. (2013). Expanding roles of protein kinase CK2 in regulating plant growth and development. J. Exp. Bot., in press.

Nakagami, H., Sugiyama, N., Ishihama, Y., and Shirasu, K. (2012). Shotguns in the front line: phosphoproteomics in plants. Plant Cell Physiol. **53:** 118–124.

Nakagami, H., Sugiyama, N., Mochida, K., Daudi, A., Yoshida, Y., Toyoda, T., Tomita, M., Ishihama, Y., and Shirasu, K. (2010). Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. Plant Physiol. **153:** 1161–1174.

Newman, R.H., et al. (2013). Construction of human activity-based phosphorylation networks. Mol. Syst. Biol. **9:** 655.

Niittylä, T., Fuglsang, A.T., Palmgren, M.G., Frommer, W.B., and Schulze, W.X. (2007). Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of Arabidopsis. Mol. Cell. Proteomics **6:** 1711–1726.

Nühse, T.S., Bottrill, A.R., Jones, A.M., and Peck, S.C. (2007). Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. Plant J. **51:** 931–940.

Nühse, T.S., Stensballe, A., Jensen, O.N., and Peck, S.C. (2004). Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. Plant Cell **16:** 2394–2405.

Oh, M.H., Clouse, S.D., and Huber, S.C. (2012a). Tyrosine phosphorylation of the BRI1 receptor kinase occurs via a post-translational modification and is activated by the juxtamembrane domain. Front. Plant Sci. **3:** 175.

Oh, M.H., Wang, X., Clouse, S.D., and Huber, S.C. (2012b). Deactivation of the Arabidopsis BRASSINOSTEROID INSENSITIVE 1 (BRI1) receptor kinase by autophosphorylation within the glycine-rich loop. Proc. Natl. Acad. Sci. USA **109:** 327–332.

Oh, M.H., Wang, X., Kota, U., Goshe, M.B., Clouse, S.D., and Huber, S.C. (2009). Tyrosine phosphorylation of the BRI1 receptor kinase emerges as a component of brassinosteroid signaling in Arabidopsis. Proc. Natl. Acad. Sci. USA **106:** 658–663.

Oh, M.H., Wu, X., Kim, H.S., Harper, J.F., Zielinski, R.E., Clouse, S.D., and Huber, S.C. (2012c). CDPKs are dual-specificity protein kinases and tyrosine autophosphorylation attenuates kinase activity. FEBS Lett. **586:** 4070–4075.

Park, H.C., Song, E.H., Nguyen, X.C., Lee, K., Kim, K.E., Kim, H.S., Lee, S.M., Kim, S.H., Bae, D.W., Yun, D.J., and Chung, W.S. (2011). Arabidopsis MAP kinase phosphatase 1 is phosphorylated and activated by its substrate AtMPK6. Plant Cell Rep. **30:** 1523–1531.

Pawson, T., and Scott, J.D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. Science **278:** 2075–2080.

Pearlman, S.M., Serber, Z., and Ferrell, J.E., Jr.,. (2011). A mechanism for the evolution of phosphorylation sites. Cell **147:** 934–946.

Reiland, S., Messerli, G., Baerenfaller, K., Gerrits, B., Endler, A., Grossmann, J., Gruissem, W., and Baginsky, S. (2009). Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. Plant Physiol. **150:** 889–903.

Reiland, S., Finazzi, G., Endler, A., Willig, A., Baerenfaller, K., Grossmann, J., Gerrits, B., Rutishauser, D., Gruissem, W., Rochaix, J.D., and Baginsky, S. (2011). Comparative phosphoproteome profiling reveals a function of the STN8 kinase in fine-tuning of cyclic electron flow (CEF). Proc. Natl. Acad. Sci. USA **108:** 12955–12960.

Riaño-Pachón, D.M., Kleessen, S., Neigenfind, J., Durek, P., Weber, E., Engelsberger, W.R., Walther, D., Selbig, J., Schulze, W.X., and Kersten, B. (2010). Proteome-wide survey of phosphorylation patterns affected by nuclear DNA polymorphisms in *Arabidopsis thaliana*. BMC Genomics **11:** 411.

Rudrabhatla, P., Reddy, M.M., and Rajasekharan, R. (2006). Genome-wide analysis and experimentation of plant serine/ threonine/tyrosine-specific protein kinases. Plant Mol. Biol. **60:** 293–319.

Schliebner, I., Pribil, M., Zühlke, J., Dietzmann, A., and Leister, D. (2008). A survey of chloroplast protein kinases and phosphatases in *Arabidopsis thaliana*. Curr. Genomics **9:** 184–190.

Schönberg, A., and Baginsky, S. (2012). Signal integration by chloroplast phosphorylation networks: an update. Front. Plant Sci. **3:** 256.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. **23:** 1391–1398.

Seet, B.T., Dikic, I., Zhou, M.M., and Pawson, T. (2006). Reading protein modifications with interaction domains. Nat. Rev. Mol. Cell Biol. **7:** 473–483.

Song, C., Ye, M., Liu, Z., Cheng, H., Jiang, X., Han, G., Songyang, Z., Tan, Y., Wang, H., Ren, J., Xue, Y., and Zou, H. (2012). Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Mol. Cell. Proteomics **11:** 1070–1083.

Steen, H., Küster, B., Fernandez, M., Pandey, A., and Mann, M. (2001). Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. Anal. Chem. **73:** 1440–1448.

Sugiyama, N., Nakagami, H., Mochida, K., Daudi, A., Tomita, M., Shirasu, K., and Ishihama, Y. (2008). Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. Mol. Syst. Biol. **4:** 193.

Tanz, S.K., Castleden, I., Hooper, C.M., Vacher, M., Small, I., and Millar, H.A. (2013). SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. Nucleic Acids Res. **41:** D1185–D1191.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37:** 914–939.

Türkeri, H., Schweer, J., and Link, G. (2012). Phylogenetic and functional features of the plastid transcription kinase cpCK2 from Arabidopsis signify a role of cysteinyl SH-groups in regulatory phosphorylation of plastid sigma factors. FEBS J. **279:** 395–409.

Umezawa, T., Sugiyama, N., Mizoguchi, M., Hayashi, S., Myouga, F., Yamaguchi-Shinozaki, K., Ishihama, Y., Hirayama, T., and Shinozaki, K. (2009). Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in Arabidopsis. Proc. Natl. Acad. Sci. USA **106:** 17588–17593.

Umezawa, T., Sugiyama, N., Takahashi, F., Anderson, J.C., Ishihama, Y., Peck, S.C., and Shinozaki, K. (2013). Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in *Arabidopsis thaliana*. Sci. Signal. **6:** rs8.

Wang, P., Xue, L., Batelli, G., Lee, S., Hou, Y.J., Van Oosten, M.J., Zhang, H., Tao, W.A., and Zhu, J.K. (2013). Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action. Proc. Natl. Acad. Sci. USA **110:** 11205–11210.

Wang, T., Kettenbach, A.N., Gerber, S.A., and Bailey-Kellogg, C. (2012). MMFPh: a maximal motif finder for phosphoproteomics datasets. Bioinformatics **28:** 1562–1570.

Wang, X., Bian, Y., Cheng, K., Gu, L.F., Ye, M., Zou, H., Sun, S.S., and He, J.X. (2013). A large-scale protein phosphorylation analysis reveals novel phosphorylation motifs and phosphoregulatory networks in Arabidopsis. J. Proteomics **78:** 486–498.

Wang, X., Bian, Y., Cheng, K., Zou, H., Sun, S.S., and He, J.X. (2012). A comprehensive differential proteomic study of nitrate deprivation in Arabidopsis reveals complex regulatory networks of plant nitrogen responses. J. Proteome Res. **11:** 2301–2315.

Wang, Y., Liu, Z., Cheng, H., Gao, T., Pan, Z., Yang, Q., Guo, A., and Xue, Y. (2014). EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases. Nucleic Acids Res. **42:** D496–D502.

Wang, Z., Dong, G., Singh, S., Steen, H., and Li, J. (2009). A simple and effective method for detecting phosphopeptides for phosphoproteomic analysis. J. Proteomics **72:** 831–835.

Whiteman, S.A., Serazetdinova, L., Jones, A.M., Sanders, D., Rathjen, J., Peck, S.C., and Maathuis, F.J. (2008). Identification of novel proteins and phosphorylation sites in a tonoplast enriched membrane fraction of *Arabidopsis thaliana*. Proteomics **8:** 3536–3547.

Wolschin, F., and Weckwerth, W. (2005). Combining metal oxide affinity chromatography (MOAC) and selective mass spectrometry for robust identification of in vivo protein phosphorylation sites. Plant Methods **1:** 9.

Wu, X.N., Sanchez Rodriguez, C., Pertl-Obermeyer, H., Obermeyer, G., and Schulze, W.X. (2013). Sucrose-induced receptor kinase SIRK1 regulates a plasma membrane aquaporin in Arabidopsis. Mol. Cell. Proteomics **12:** 2856–2873.

Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. (2008). GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol. Cell. Proteomics **7:** 1598–1608.

Yachie, N., Saito, R., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011). Integrative features of the yeast phosphoproteome and protein-protein interaction map. PLoS Comput. Biol. **7:** e1001064.

Yang, Z., Guo, G., Zhang, M., Liu, C.Y., Hu, Q., Lam, H., Cheng, H., Xue, Y., Li, J., and Li, N. (2013). Stable isotope metabolic labeling-based quantitative phosphoproteomic analysis of Arabidopsis mutants reveals ethylene-regulated time-dependent phosphoproteins and putative

substrates of constitutive triple response 1 kinase. Mol. Cell. Proteomics **12:** 3559–3582.

Yao, Q., Bollinger, C., Gao, J., Xu, D., and Thelen, J.J. (2012). P(3) DB: An Integrated Database for Plant Protein Phosphorylation. Front. Plant Sci. **3:** 206.

Zeng, Y., Pan, Z., Wang, L., Ding, Y., Xu, Q., Xiao, S., and Deng, X. (2014). Phosphoproteomic analysis of chromoplasts from sweet orange during fruit ripening. Physiol. Plant. **150:** 252–270.

Zhang, H., Zhou, H., Berke, L., Heck, A.J., Mohammed, S., Scheres, B., and Menke, F.L. (2013). Quantitative phosphoproteomics after auxin-stimulated lateral root induction identifies an SNX1 protein phosphorylation site required for growth. Mol. Cell. Proteomics **12:** 1158–1169.

Zulawski, M., Braginets, R., and Schulze, W.X. (2013). PhosPhAt goes kinases—searchable protein kinase target information in the plant phosphorylation site database PhosPhAt. Nucleic Acids Res. **41:** D1176–D1184.